nature ecology & evolution

Article

https://doi.org/10.1038/s41559-024-02398-5

Evolution of tissue-specific expression of ancestral genes across vertebrates and insects

Received: 2 August 2023

Accepted: 8 March 2024

Published online: 15 April 2024

Check for updates

Federica Mantica[1,2], Luis P. Iñiguez[1], Yamile Marquez[1], Jon Permanyer[1], Antonio Torres-Mendez[1], Josefa Cruz[3], Xavier Franch-Marro[3], Frank Tulenko[4], Demian Burguera[1], Stephanie Bertrand[5], Toby Doyle[6], Marcela Nouzova[7], Peter D. Currie[4,8], Fernando G. Noriega[9,10], Hector Escriva[5], Maria Ina Arnone[11], Caroline B. Albertin[12], Karl R. Wotton[6], Isabel Almudi[13], David Martin[3] & Manuel Irimia[1,2,14]

Regulation of gene expression is arguably the main mechanism underlying the phenotypic diversity of tissues within and between species. Here we assembled an extensive transcriptomic dataset covering 8 tissues across 20 bilaterian species and performed analyses using a symmetric phylogeny that allowed the combined and parallel investigation of gene expression evolution between vertebrates and insects. We specifically focused on widely conserved ancestral genes, identifying strong cores of pan-bilaterian tissue-specific genes and even larger groups that diverged to define vertebrate and insect tissues. Systematic inferences of tissue-specificity gains and losses show that nearly half of all ancestral genes have been recruited into tissue-specific transcriptomes. This occurred during both ancient and, especially, recent bilaterian evolution, with several gains being associated with the emergence of unique phenotypes (for example, novel cell types). Such pervasive evolution of tissue specificity was linked to gene duplication coupled with expression specialization of one of the copies, revealing an unappreciated prolonged effect of whole-genome duplications on recent vertebrate evolution.

Fossil records reconstruct the image of the last common ancestor (LCA) of all bilaterian animals as a small, marine creature[1] crawling on the sea-floor approximately 700 million years ago (Ma)[1]. Despite its apparent simplicity, this ancestral organism already possessed an ancestral form of the main tissue types that are homologous across extant bilaterian species, including a nervous system, skeletal muscle, female and male gonads, a gut and an excretory system[2]. How did this ancient organism specify such a great variety of biological structures? Since all its cells

[1]Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Universitat Pompeu Fabra, Barcelona, Spain. [3]Institute of Evolutionary Biology (IBE, CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia, Spain. [4]Australian Regenerative Medicine Institute, Monash University, Clayton, Victoria, Australia. [5]Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins; BIOM, Banyuls-sur-Mer, France. [6]Centre for Ecology and Conservation, University of Exeter, Penryn, UK. [7]Institute of Parasitology, CAS, České Budějovice, Czech Republic. [8]EMBL Australia; Victorian Node, Monash University, Clayton, Victoria, Australia. [9]Biology and BSI, Florida International University, Miami, FL, USA. [10]Department of Parasitology, University of South Bohemia, České Budějovice, Czech Republic. [11]Stazione Zoologica Anton Dohrn, Napoli, Italy. [12]Eugene Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, MA, USA. [13]Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Barcelona, Spain. [14]ICREA, Barcelona, Spain. e-mail: mirimia@gmail.com

shared the same genome, gene expression regulation was likely key for the generation of unique transcriptomes across these ancestral tissue types and consequently, for the emergence of their distinctive biological functions.

The bilaterian ancestor gave rise to the vast majority of extant animals, where the original body plan and tissues have been greatly diversified and modified. Determinants of animal evolution include changes in gene complements (that is, gene gains/losses and gene duplications)[3,4], divergence of protein-coding sequences[5] and regulatory changes in gene expression[6]. In fact, as the generation of tissue-specific transcriptomes is key for defining distinct tissue types (that is, intra-species diversity), their evolutionary remodelling is arguably the most crucial determinant of phenotypic variation among species (that is, inter-species diversity)[7,8]. Importantly, major transcriptome remodelling often involves conserved genes[9]. Examples of such occurrences underlie important phenotypic novelties in key bilaterian lineages: the vertebrate endocrine pancreas emerged following the recruitment of ancestrally neural-specific genes[10], and it has been suggested that insect wings evolved upon co-option of expression of ancient genes originally involved in gill specification and proximal leg segments[11–13]. Still, even if remarkable cases linked to biological novelties have been identified, the specific role that the evolution of expression of ancestral genes played in shaping homologous, yet often highly divergent, tissues between distant bilaterian lineages has never been thoroughly assessed.

Here we studied the evolution of tissue-specific transcriptomes on the basis of an extensive RNA sequencing (RNA-seq) dataset covering 8 tissue types from 20 bilaterian species, including novel data for 15 of them. Compared with previous studies of transcriptome evolution, mostly focused on mammals, this dataset extends the phylogenetic coverage beyond vertebrates and provides presumably the first comparative framework of this scale for insects. Vertebrates and insects were selected as focus clades because they include highly accessible organisms with a relatively similar body plan compared with other lineages (for example, most lophotrochozoans). However, they also reached opposite evolutionary solutions in terms of structural organization (for example, dorsal/ventral positioning of the spinal/nerve cords in vertebrates and insects, respectively[14]), and show different molecular and genomic evolutionary rates[15,16]. We selected a symmetric phylogeny for both clades, composed of two branches with pairs of vertebrate and insect species in equivalent phylogenetic position and with similar evolutionary distances. This structure allowed us to perform sound ancestral inferences as well as to uncover parallel, convergent and divergent evolutionary trajectories in comparable evolutionary/phylogenetic nodes of these early diverging bilaterian lineages. First, we characterized global gene expression patterns and reconstructed ancestral bilaterian tissue-specific modules that are still widely conserved across extant species. Second, we systematically inferred gains and losses of tissue-specific expression throughout our selected bilaterian phylogeny. Lastly, we characterized these inferred tissue-specificity gains from the mechanistic and functional perspective. Overall, our work sheds light on the highly plastic nature of deeply conserved genes in terms of tissue-specific expression patterns, which we find to be tightly linked to gene duplication, specialization and the emergence of unique tissue-related phenotypes.

Results

Global patterns of gene expression across bilaterian tissues

To reliably investigate the evolution of tissue-specific (TS) transcriptomes in two key bilaterian lineages, we selected 20 representative species (8 gnathostome vertebrates, 8 insects and 2 pairs of relative outgroups) evenly divided into two monophyletic branches with specular phylogenetic structures (Fig. 1a). After correcting for broken/chimaeric genes and enriching the annotations of the majority of the species (see Supplementary Methods, Extended Data Fig. 1a–d and Supplementary Data 1), we derived gene orthology relationships among all of them and isolated 7,178 bilaterian-conserved gene orthogroups, which were unambiguously present in the bilaterian LCA (that is, conserved in at least 12/20 species) (see Supplementary Methods, Extended Data Fig. 1e,f and Supplementary Data 2). In addition, as several comparisons required the selection of one representative orthologue per species, we generated the most suitable one-to-one orthogroups for each set of analyses (that is, best-ancestral and best-TS orthogroups; see Extended Data Fig. 2 and Supplementary Methods) by applying distinct filters to the original bilaterian-conserved orthogroups. We then assembled an extensive bulk RNA-seq dataset covering up to 8 tissues in all species (Fig. 1a and Supplementary Data 3). Notably, while some of the included tissue types (neural, testis, ovaries, muscle and excretory system) had been considered in previous studies of gene expression evolution among bony vertebrate species[17–20], others (digestive tract, epidermis and adipose) have never been analysed in such a context. In total, we generated 89 RNA-seq samples across 15 species, which we combined with publicly available data into a final dataset of 346 RNA-seq metasamples (see Methods for metasamples definition), including up to 3 metasamples for each tissue and species (Fig. 1a and Supplementary Fig. 1).

In our dataset, the first two components of a principal component analysis (PCA) showed a clear distinction between metasamples from vertebrates and non-vertebrates (including all outgroups), independent of their tissue identity (Fig. 1b; see Supplementary Methods for normalization procedure). The same pattern is observed when controlling for potential batch effects (Supplementary Fig. 2a–f) or using alternative strategies for the selection and expression quantification of the best-ancestral orthogroups (Supplementary Fig. 2g–j). Thus, this separation probably emerged because of some intrinsic differences between vertebrates and insects, which we further investigated by performing clade-specific PCAs (see Methods). The first two components of the vertebrate PCA outlined groups of metasamples of the same tissue origin (specially for neural and testis; Supplementary Fig. 3a,b), confirming the pattern characterized by previous studies[17,20,21] and validating the idea of a conserved, tissue-related transcriptomic signature that prevails over the species identity. On the other hand, the first two components of the insect PCA were dominated by species clustering (Supplementary Fig. 3c), with successive components separating species according to their evolutionary distances (Supplementary Fig. 3d). Altogether, these results thus suggest faster evolutionary rates within and between insects compared with vertebrates, as reflected by the vertebrate/non-vertebrate metasamples separation in the first two components of the bilaterian PCA (Fig. 1b). Nevertheless, some subsequent components of this PCA significantly separated groups of metasamples on the basis of their tissue of origin, starting with neural and testis tissues (Extended Data Fig. 3a–c). Thus, we performed a *z*-score transformation of gene expression values within species (see Supplementary Methods) to compare their relative expression profiles across tissues. Upon this transformation, we obtained clusters largely corresponding to tissue groups (Extended Data Fig. 3d), which suggested at least partial conservation of ancestral bilaterian tissue-specific expression modules.

Reconstructing ancestral tissue-specific expression modules

To characterize ancestral tissue-specific expression modules that are still widely conserved across extant bilaterians, we implemented a strategy based on a sparse partial least square discriminant analysis (sPLS-DA)[22], which allowed us to isolate the orthogroups with the most distinctive expression profiles in each tissue (compared with the others) across all species (see Methods, Fig. 2a and Extended Data Fig. 4a–f). The sPLS-DA approach was set up to produce eight sets (one per tissue type) of ancestral orthogroups with conserved tissue-specific expression, which overall comprised 506 (~7%) of all bilaterian-conserved orthogroups (Fig. 2b and Supplementary Data 4). A PCA performed on

https://doi.org/10.1038/s41559-024-02398-5

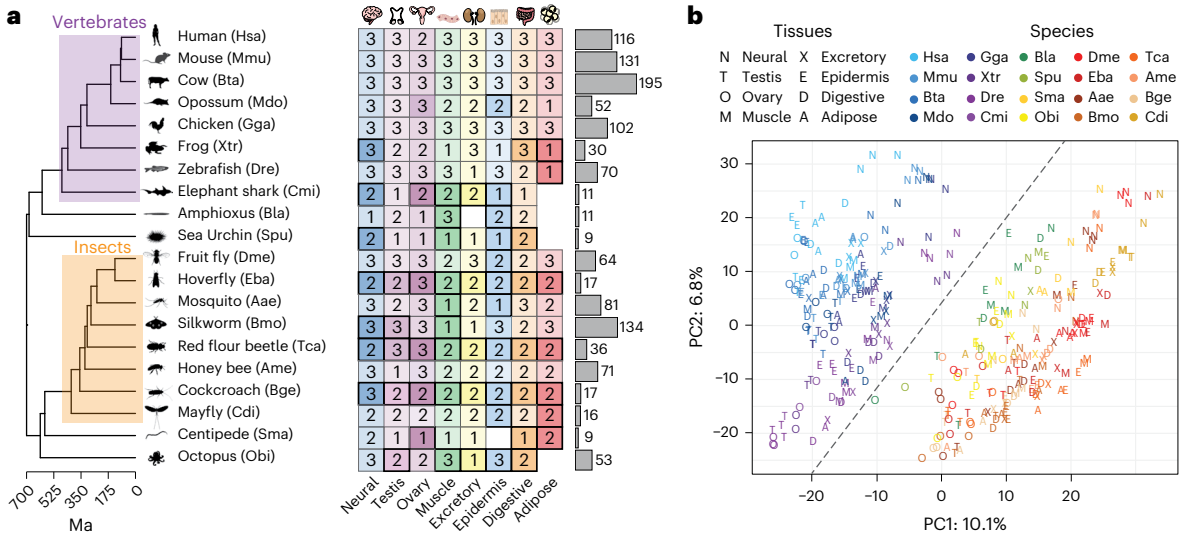


Fig. 1 | Dataset overview and global patterns of gene expression across bilaterian tissues. a, RNA-seq dataset overview. Left: phylogenetic tree including the common names and scientific acronyms of the 20 bilaterian species considered in this study. Evolutionary distances were derived from ref. 69. Animal silhouettes were downloaded from http://phylopic.org/, with credits to Sarah Werning for the opossum (https://creativecommons.org/licenses/by/3.0/), Soledad Miranda-Rottman for the frog (https://creativecommons.org/licenses/by/3.0/), Tony Ayling and Milton Tan for the elephant shark (https://creativecommons.org/licenses/by-sa/3.0/), Harold N. Eyster for the sea urchin (https://creativecommons.org/licenses/by/3.0/), Gareth Monger for the hoverfly (https://creativecommons.org/licenses/by/3.0/) and Birgit Lang for the centipede (https://creativecommons.org/licenses/by/3.0/). Tissue icons were created by Queralt Tolosa. Middle: scheme of RNA-seq metasamples. The number of metasamples for each species (rows) and tissue (columns) is reported. The cell colour corresponds to the tissue, while its intensity and border thickness distinguishes between cases where at least one RNA-seq sample has been generated for this project (full colour, thick borders) from cases where all the included samples are publicly available (transparent colour, thin borders). Right: barplot with the total number of processed RNA-seq samples per species. **b**, Coordinates of the first (PC1; *x* axis) and second (PC2; *y* axis) principal components from a PCA performed on the best-ancestral orthogroups normalized expression matrix (see Methods). Only the 2,436 best-ancestral orthogroups conserved in all species were considered. Tissue identity is represented by letters and species by colours. The percentage of variance explained by each PC is reported on the relative axis. The dashed line separates vertebrate from non-vertebrate metasamples.

these ancestral orthogroups (see Methods and Supplementary Fig. 4a), or subsets of these orthogroups of the same size across tissues (Supplementary Fig. 4b), showed aggregation by tissue type, with a clear separation between all neural and non-neural metasamples along the first principal component.

We next investigated in detail the orthogroups belonging to the neural and testis modules whose expression profiles are shown in Fig. 2c,d (see Extended Data Fig. 4g–l for other modules). The neural module presented strong over-representation of gene ontology (GO) categories related to synaptic transmission, neuronal morphology and other associated terms (Fig. 2e), reflecting the high expression conservation of the specialized neuronal gene complement across eumetazoan nervous systems[23,24]. The testis module showed significant enrichments for cilium and cytoskeleton-related functions (Fig. 2f), probably determined by the axoneme, a highly conserved microtubule-based structure located at the core of most bilaterian spermatozoa flagella and indispensable for their mobility[25]. The relevant role that these ancestral orthogroups probably play in the respective tissue is supported by their significantly greater association with validated neural- or testis-related phenotypes either in mammals or fly, compared with all bilaterian-conserved orthogroups (Fig. 2g,h and Supplementary Data 5; $P < 1 \times 10^{-5}$ for both tissues, Fisher's exact tests).

In addition to the neural and testis modules, all other sets exhibited GO enrichments coherent with the deep-rooted functions of each tissue, and comparable between the human (Fig. 2i and Supplementary Data 6) and fruit fly-based GO annotations (Supplementary Data 7). For instance, genes in the ancestral ovary module comprised several key meiotic genes and were enriched in cell cycle and DNA-replication/repair functions (Supplementary Data 4, 6 and 7). Some examples included *CCNB2*, a cyclin necessary for timely oocyte maturation and correct metaphase-to-anaphase transition in mice[26,27]; *MOS*, a serine–threonine kinase that mediates metaphase II arrest during meiosis and whose deletion causes human female infertility[28]; and *CPEB*, a protein involved in the regulation of translation before fertilization[29]. As another example, the most significant GO categories for the excretory system module, mainly ion transport and amino acid metabolism, reflected the basic shared functions of ultrafiltration-based excretory systems[30]. Moreover, even in those tissues in which the homology status of the specific cellular components is more ambiguous/complex (for example, epidermis, digestive system, adipose), we still obtained a few significant enrichments linked to core molecular programmes underlying fundamental functions of each tissue type (Fig. 2i).

Finally, we investigated which transcription factors (TFs) might have regulated these ancestral modules since the bilaterian LCA. We specifically tested whether the TFs included in each module presented a significant over-representation of predicted binding sites in the regulatory regions of all the other genes in the same ancestral set (see Supplementary Methods). We obtained significant results for multiple TFs comprising several known master regulators of the respective tissues, such as *PAX4/6* (ref. 31) or *FEZF1* (ref. 32) in neural, *MEF2A-D*[33] in muscle and *GRHL1/2* (ref. 34) in epidermis. Importantly, over-representation of their binding sites was observed both at the module level (Fig. 2j) and within each studied species separately (Supplementary Fig. 4c).

Pervasive evolution of tissue specificity of ancestral genes

To study the evolution of tissue specificity throughout our entire phylogeny, we next used the Tau metric to define genes with tissue-specific expression profiles in all extant species[35]. The overall proportion of tissue-specific genes (Tau ≥ 0.75) in each species was lower for bilaterian-conserved genes (that is, belonging to the 7,178 bilaterian-conserved orthogroups) compared with all genes (Fig. 3a and Supplementary Fig. 5). This was partially expected, as highly conserved genes are usually associated with greater pleiotropic roles. We assigned each bilaterian-conserved, tissue-specific gene to the tissue(s) with the

https://doi.org/10.1038/s41559-024-02398-5

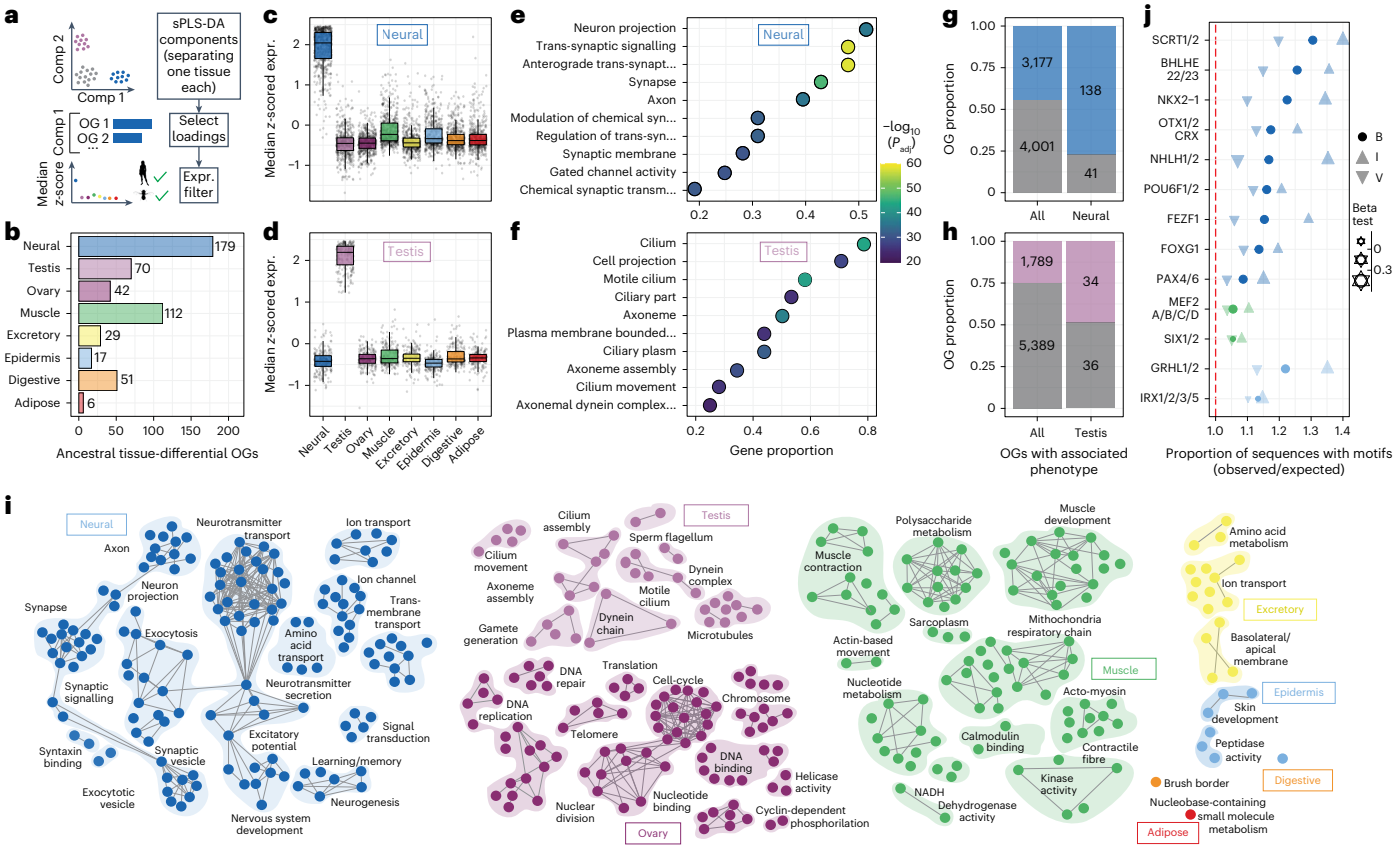


Fig. 2 | Reconstruction of ancestral bilaterian tissue-specific expression modules. a, Scheme depicting the procedure for the definition of the ancestral tissue-specific modules. Animal silhouettes were downloaded from http://phylopic.org/. **b**, Number of best-ancestral orthogroups (OGs) included in each ancestral tissue-specific module. **c**,**d**, Expression profiles across tissues of best-ancestral orthogroups in the ancestral neural- (**c**; $n = 179$) and testis- (**d**; $n = 70$) specific modules. Expression values were first *z*-scored by species and each dot represents the median expression among vertebrates, insects or outgroups. In boxplots: the centre line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend to the lowest and highest points to a limit of 1.5× the interquartile range from the closest hinge. **e**,**f**, Top 10 most significantly enriched GO categories in the ancestral neural- (**e**) and testis- (**f**) specific modules. Only GOs containing at least 5 orthogroups in the tested set were considered. **g**,**h**, Proportion of all bilaterian-conserved orthogroups (left) and ancestral neural- (**g**) or testis-specific (**h**) modules (right) associated with experimentally validated phenotypes (in the respective tissue) in mammals and/or fruit fly. **i**, Representation of GO networks of significantly enriched categories for all ancestral tissue-specific modules, where only categories containing at least 5 orthogroups in the tested set were considered (see Methods). Each node represents a GO category. **j**, TFs included within each tissue-specific module whose known binding motifs are significantly over-represented in the regulatory regions of the genes in the corresponding module (see Supplementary Methods). Each TF was tested (Fisher's exact and regression tests) on all sequences (B, bilaterian; V, only vertebrate; I, only insect) within the module. TFs in each tissue are ordered by the ratio of the proportion of sequences with at least one predicted binding site in the tested module (observed) compared to the proportion in all other bilaterian-conserved genes (expected). The size of each dot reflects the beta from the regression test in the corresponding group and tissue colours refer to **b**.

highest relative expression (see Methods, Fig. 3b, Extended Data Fig. 2c and Supplementary Fig. 6), providing a comprehensive characterization of their tissue specificity across all species and tissues (Fig. 3c). Neural- and testis-specific genes were the most abundant throughout our phylogeny, followed by genes with restricted expression in two different tissues (Fig. 3d). Overall, the number of tissue-specific genes was significantly higher among vertebrate species (Fig. 3e; $P = 2 \times 10^{-4}$; Wilcoxon rank-sum test), probably because of the two whole-genome duplications (WGDs) at the base of vertebrates (see below).

We then set out to investigate the conservation of the identified tissue-specific profiles. Remarkably, we found that these profiles were overall poorly conserved. For instance, for the orthogroups that have at least one gene that is tissue specific in mouse, only a median of 6 out of the other 19 species had at least one orthologue with Tau ≥ 0.75 (with any associated tissue) and this number merely increased to 9 for Tau ≥ 0.5 (Fig. 3f). This pattern was consistent across all species (Fig. 3g and Supplementary Fig. 7), suggesting that tissue specificity is highly dynamic and that a high proportion of these tissue-specific expression profiles may have a recent evolutionary origin. In fact, when considering individual species alone, only between 4% and 15% of bilaterian-conserved orthogroups are tissue specific in each of them (barplot in Fig. 3h); however, when all species are taken together, 47% of the bilaterian-conserved orthogroups contain at least one tissue-specific gene (line plot in Fig. 3h). In other words, the orthogroups containing tissue-specific genes are widely non-overlapping among species. As expected, orthogroups that are tissue specific in at least one studied species presented a significantly higher proportion of gene duplications compared with orthogroups that are never tissue specific ($P = 2 \times 10^{-4}$, Fisher's exact test), which in turn were strongly enriched for housekeeping functions such as RNA processing/binding and translation (extra boxes in Fig. 3h and Supplementary Data 8).

Systematic inferences of tissue-specificity gains and losses

Next, to investigate the dynamics of tissue-specificity evolution in finer detail throughout the entire phylogeny, we adopted a parsimony-based approach to perform a systematic phylogenetic inference of tissue-specificity gains and losses in each tissue (see Methods, Extended Data Fig. 5 and Supplementary Data 9; for the rationale behind our chosen

https://doi.org/10.1038/s41559-024-02398-5

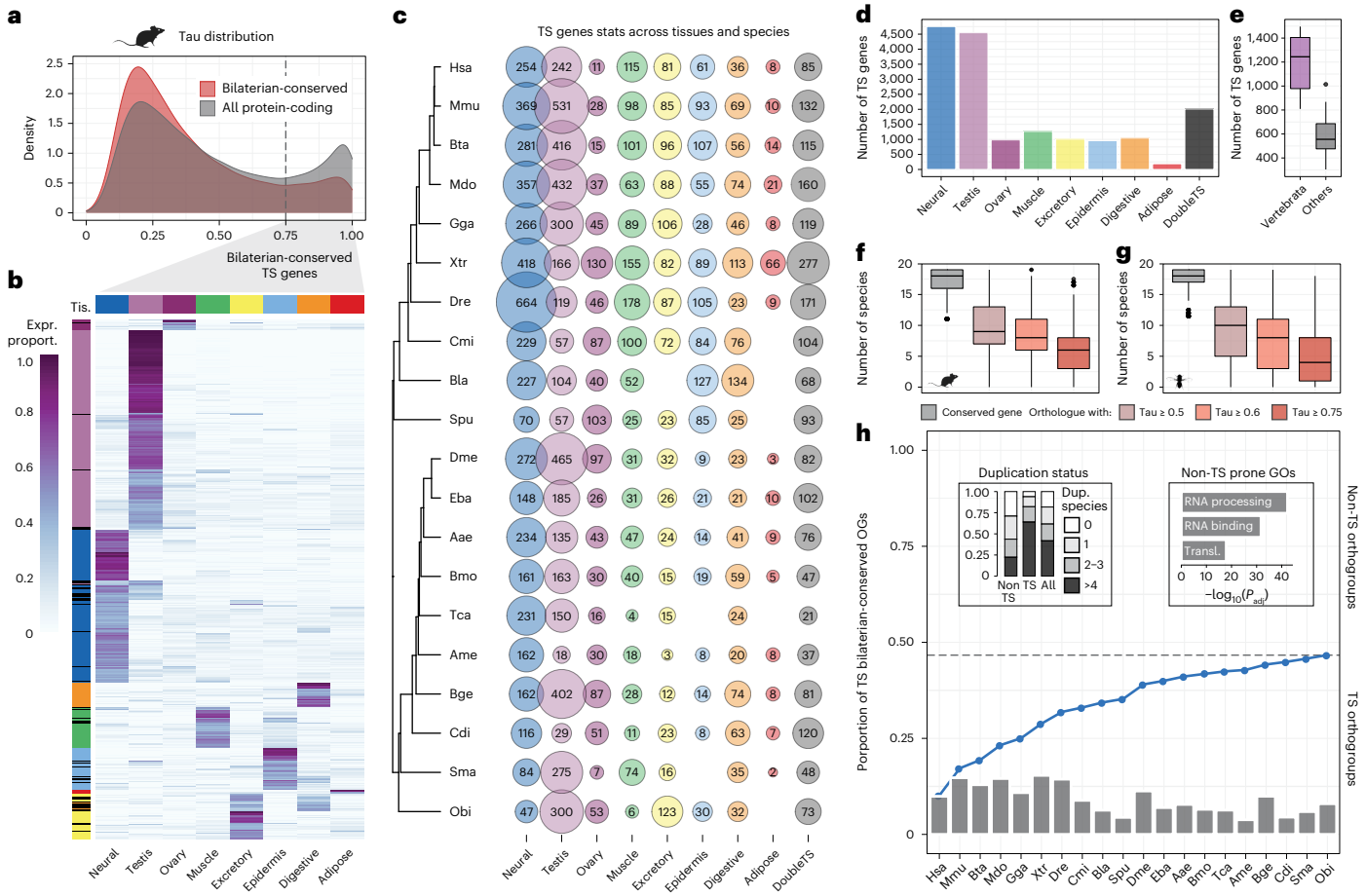


Fig. 3 | Tissue-specificity patterns across species reveal low conservation of tissue-specific expression profiles. a, Tau distributions of all (grey) or bilaterian-conserved (red) mouse protein-coding genes passing the expression cut-off. **b**, Heatmap showing the clustering of mouse bilaterian-conserved, tissue-specific genes (rows) based on their expression proportion (tissue_expr/all_tissue_expr) across tissues (columns). The heatmap was generated using pheatmap in R with default parameters, and the complete dendrogram is shown in Supplementary Fig. 6. Black indicates double tissue specificity. **c**,**d**, Number of bilaterian-conserved, tissue-specific genes across all species (rows) and tissues (columns) (**c**) and collapsed by tissue (**d**). **e**, Distribution of the number of bilaterian-conserved, tissue-specific genes in vertebrates ($n = 8$) versus all other species ($n = 12$) ($P = 2 \times 10^{-4}$, two-sided Wilcoxon rank-sum test). In boxplots: the centre line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend to the lowest and highest points to a limit of 1.5× the interquartile range from the closest hinge. Data points beyond the end of the whiskers are plotted individually. **f**,**g**, Distribution of the number of species in which mouse (**f**) or fruit fly (**g**) bilaterian-conserved, tissue-specific genes ($n = 1{,}415$ and $n = 1{,}014$, respectively) have at least one orthologue (grey) and this orthologue has a Tau value higher than a specific cut-off (0.5, 0.6 and 0.75; other shades) (see **e** for description of boxplot features). **h**, Barplot: proportion of bilaterian-conserved orthogroups including at least one tissue-specific gene in each given species. Line plot: cumulative distribution of the proportion of unique bilaterian-conserved orthogroups containing at least one tissue-specific gene across species. All genes included in the bilaterian-conserved orthogroups were considered for this analysis. The dashed line marks the total proportion. The two boxes include information on the duplication status of the non-TS and TS orthogroups (left) and the top GO enrichments for the non-TS orthogroups (right). Animal silhouettes (**a**,**f**,**g**) were downloaded from http://phylopic.org/.

procedure, the comparison with other inference methods and their limitations for our dataset, see Supplementary Discussion). By definition, given the lack of non-bilaterian outgroups in our phylogeny, we could only infer tissue-specificity gains and losses posterior to the last bilaterian ancestor. Thus, we focused only on the patterns of gain/loss on the phylogenetically equivalent nodes along the two main branches and the tips leading to the extant species (Fig. 4a). These inferences, as well as the underlying Tau values, were highly robust to the use of alternative combinations of the original RNA-seq samples (see Supplementary Methods), computed either after averaging the expression of all available samples for each tissue (Supplementary Fig. 8) or after randomizing the samples of each tissue across the relative metasamples (Supplementary Fig. 9). Moreover, since we observed that gene expression divergence is subjected to stabilizing selection in all tissues along both phylogenetic branches and thus mainly evolves following an Ornstein–Uhlenbeck (OU) curve (Supplementary Fig. 10), we used an OU-based approach to orthogonally validate our inferred tissue-specificity gains and losses, finding a good overall agreement (see Extended Data Fig. 6, and Supplementary Methods and Discussion).

According to our inference methodology, the Vertebrata ancestor presents the highest overall number of inferred tissue-specificity gains among all ancestral nodes (Fig. 4a), even if some of the most important gain waves were observed for individual species (for example, in fruit fly testis and frog ovary, with 198 and 109 gains, respectively; Fig. 4a and Extended Data Fig. 7a). In this regard, we found that some genomic features were positively correlated with the number of species-specific, tissue-specific gains, but none of them reached statistical significance (Supplementary Fig. 11a–e).

Comparison of the proportions of tissue-specificity gains and losses within each node and species shows that testis had the highest turnover, as it presented the greatest proportion of gains and/or losses in 30/39 (77%) of nodes/species (Fig. 4b). On the contrary, neural gains are mainly prevalent in the most ancestral nodes on both branches (that is, Euteleostomi/Neoptera or older), but while they seem to have little

https://doi.org/10.1038/s41559-024-02398-5

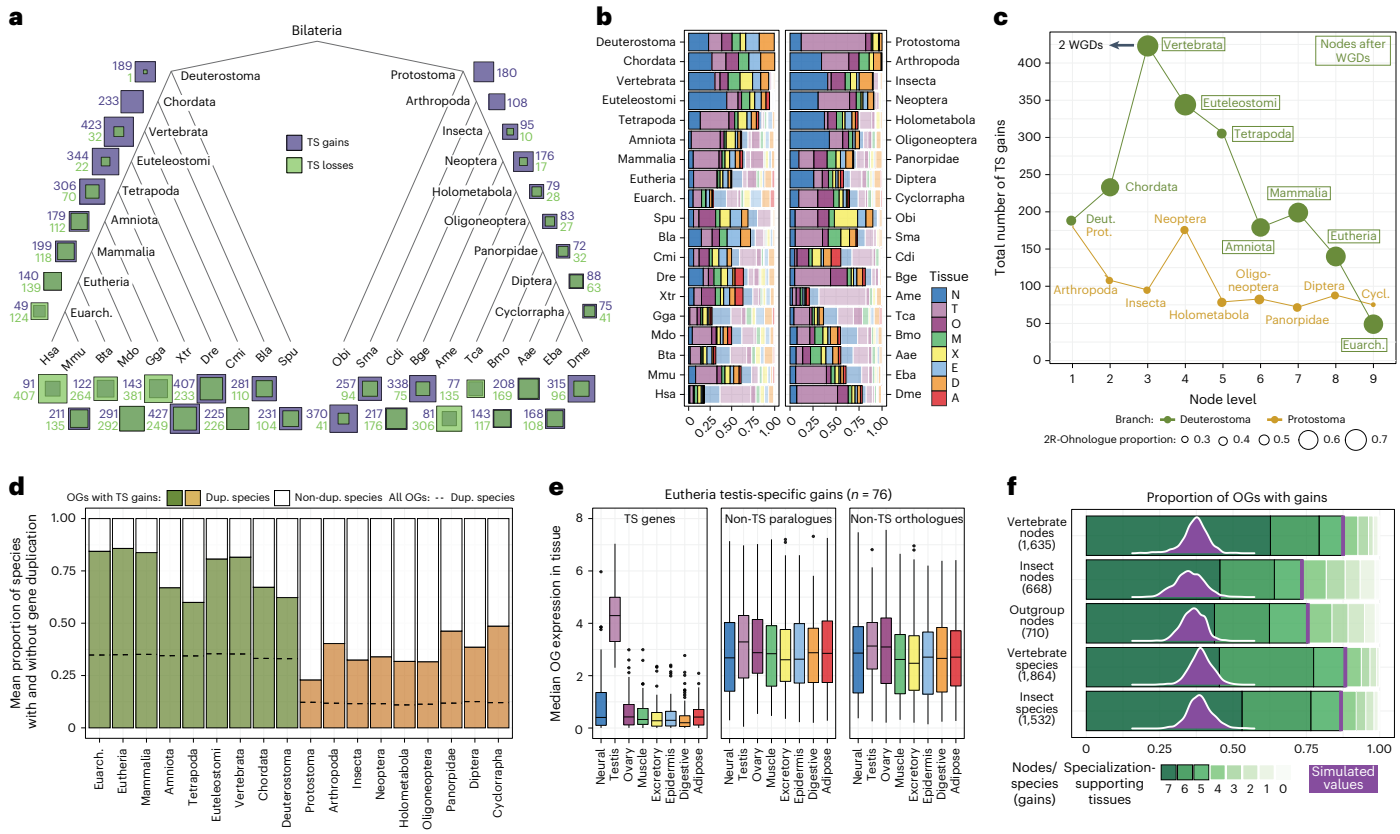


Fig. 4 | Tissue-specificity gains are associated with gene duplication and specialization. a, Total numbers of tissue-specificity gains and losses across all nodes and species. **b**, Relative proportions of tissue-specificity gains and losses across tissues within each node and species. Full/transparent shades of tissue colours represent gains/losses. **c**, Total number of tissue-specificity gains across nodes on each phylogenetic branch. The size of the dots represents the proportion of orthogroups including 2R-ohnologues (that is, paralogues originated by the two rounds of vertebrate WGDs) in each gain group. **d**, Average proportions of duplicated and non-duplicated species among the species with tissue-specific expression in the orthogroups that gain tissue specificity in each node. The background dashed line represents the expected proportion based on all bilaterian-conserved orthogroups for the same sets of species (that is, descendant species for that node). **e**, Median gene expression across tissues for bilaterian-conserved orthogroups with testis-specific gains in Eutheria (76 orthogroups). Left: testis best-TS orthologues in eutherians (3 species). Middle: eutherian non-testis-specific paralogues. Right: testis best-TS orthologues in non-eutherian species (17 species). In boxplots: the centre line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend to the lowest and highest points to a limit of 1.5× the interquartile range from the closest hinge. Data points beyond the end of the whiskers are plotted individually. **f**, For each set of tissue-specificity gains, distribution of the number of tissues (in which the gene is not tissue specific) where the median expression of the species without tissue specificity is higher than in the set of species with tissue specificity (from 0 to 7, 'specialization-supporting tissues'). The purple distribution represents the proportion of gains with specialization-supporting tissues ≥5 coming from 100 randomizations of the tissue-specificity labels within the respective best-TS orthogroups (see Extended Data Fig. 7d,e for full data). N, neural; T, testis; O, ovary; M, muscle; X, excretory; E, epidermis; D, digestive; A, adipose; Euarch, Euarchontoglires; Cycl, Cyclorrapha; Deut, Deuterostoma; Prot, Protostoma.

impact in later vertebrate evolution, they still dominate the gain landscape in several more recent insect nodes (for example, Holometabola, Oligoneoptera and Diptera). Finally, as opposed to the most ancestral nodes, the tissue-specific transcriptome of few recent nodes and of the majority of single species is predominantly shaped by losses of tissue specificity rather than by gains (that is, we observed an average of 11% of losses on the total number of inferences for Tetrapoda/Holometabola and more ancient nodes, compared with 42% for more recent nodes and single species). While this result is expected, as by definition we infer losses of tissue specificity only after a tissue-specificity gain has been identified in more ancestral nodes, it is still relevant to highlight the large plasticity of the expression profiles of ancestral genes during recent vertebrate and insect evolution.

We then evaluated enrichments of orthogroups with tissue-specificity gains across gene families, obtaining similar results upon definition of gene families based either on the human (Supplementary Fig. 12a) or fruit fly (Supplementary Fig. 12b) annotation (see Supplementary Methods). On one hand, we found significant over-representation of tissue-specificity gains for many families of membrane proteins, including several types of transporter and receptor (for example, ion channels, EGF transporters, G protein-coupled receptors and so on) (Supplementary Fig. 12a,b). On the other hand, families of housekeeping genes (for example, ribosomal RNA, histone proteins) or enzymes devoted to basic cellular metabolic processes (for example, transfer RNA synthetases, acetyl-transferases) were significantly depleted for tissue-specificity gains, in line with the results reported in Fig. 3h and Supplementary Data 8.

Tissue specificity is tied to duplication and specialization

We then compared the total numbers of tissue-specificity gains between phylogenetically equivalent nodes on the two branches (Fig. 4c). The gain signal reached the overall maximum in the Vertebrata ancestor, consistent with a strong impact of the two rounds of WGD at the origin of this group. Strikingly, although this effect progressively decreased, the fraction of gains remained high throughout all subsequent vertebrate nodes, in clear contrast with the relative flat signal observed on the insect side (except for the Neoptera ancestor, in line with a burst of gene duplication events[36]). Moreover, gains in both the Vertebrata and

subsequent nodes, as well as in the species-specific branches, showed a much higher proportion of orthogroups involving paralogues derived from the vertebrate WGDs (2R-ohnologues) compared with phylogenetically equivalent Insecta nodes and species (Fig. 4c and Extended Data Fig. 7b). Altogether, this suggests the existence of a previously unappreciated prolonged evolutionary impact of vertebrate WGDs on the rewiring of tissue-specific transcriptomes.

Remarkably, the association between the gain of tissue specificity and gene duplication extended beyond the vertebrate WGDs. For all nodes and extant species, we found that orthogroups with inferred tissue specificity had a higher proportion of duplicates compared with the corresponding background (Fig. 4d and Extended Data Fig. 7c). Importantly, while the chances of having a tissue-specific gene are expected to increase with the number of paralogues in an orthogroup, randomizations showed that this effect cannot explain the observed association between gene duplication and tissue specificity (Extended Data Fig. 8a).

Moreover, we found evidence that the acquisition of tissue-specific expression occurred to a large extent through the process of specialization[37], where the specialized paralogue reduces its expression in most tissues, while (1) the other paralogue(s) and (2) their orthologues in other species conserve the ancestral, broader expression pattern (see scheme in Extended Data Fig. 8b). For example, under this model, the Eutheria testis-specific genes are expected to have lower expression across the other tissues (that is, all tissues but testis) compared with (1) their non-testis-specific paralogues in eutherians and (2) their orthologues in non-eutherians, as readily observed in our dataset (Fig. 4e). Both these patterns were consistently observed across all nodes and species (Extended Data Fig. 8c,d and Supplementary Dataset), which prompted us to systematically evaluate the incidence of specialization events. We used as a metric the number of 'specialization-supporting' tissues, corresponding to all those tissues (excluding the tissue with tissue specificity) where the expression of a given orthologue is lower in the set of species with tissue specificity compared with those without. We plotted the total number of specialization-supporting tissues (from 0 to 7) for tissue-specificity gains across different groups of nodes and species (Fig. 4f), showing that, in all cases, specialization events (that is, number of specialization-supporting tissues ≥5) occurred more extensively than expected by chance (Fig. 4f and Extended Data Fig. 7d,e).

Tissue-specificity gains and phenotypic evolution

We also identified 156 bilaterian-conserved orthogroups that acquired unique but distinct tissue specificity on the vertebrate and insect sides, thus fulfilling their functional potential in divergent contexts (Extended Data Fig. 9a). The most frequent pairs of tissues among which these parallel tissue-specificity gains occurred were neural and testis together with testis and ovary (Extended Data Fig. 9b), in agreement with the compartments among which expression shifts are more likely to occur also within vertebrates[20]. In addition to these divergent/parallel tissue-specificity gains, we also characterized independent convergent acquisitions of the same tissue-specific expression profiles in both the vertebrate and the insect sides (Fig. 5a). Such convergent gains were most abundant in testis, probably due to the faster turnover of tissue specificity this tissue experienced in both clades (Fig. 4b). One exemplary case is represented by *TESMIN* and *tomb* (Fig. 5b). These are paralogues of the ancestral *LIN54/mip120* gene that independently originated in the vertebrate and insect lineages, convergently acquired testis-specific expression in amniotes and the fruit fly, respectively, and whose importance for testis development and function is proven by spermatogenesis disruption upon gene perturbation both in mouse[38] and fruit fly[39].

Next, we aimed to functionally characterize the tissue-specificity gains in each node and species (Supplementary Data 10). We found that a few gene functions were significantly and repeatedly enriched in multiple nodes/species on both phylogenetic branches (Fig. 5c). GO categories such as DNA binding and cation transmembrane transport were over-represented throughout all tissue types, but we also identified a few functions specifically enriched across somatic organ tissues (for example, plasma membrane region, consistent with the gene family analyses (Supplementary Fig. 11)) or reproductive ones (mainly related to meiotic division). Moreover, GO enrichments performed with insect-specific GO annotations (see Methods) showed repeated enrichments for orthogroups involved in flight and cuticle formation in muscle-specific and epidermis-specific gains, respectively, throughout insect evolution (Supplementary Data 11 and 12). In contrast, each tissue presented categories exclusively enriched across gains in a single node/species (Fig. 5d and Supplementary Data 13), several of which could be linked to the concurrent emergence of novel phenotypic traits. For instance, only vertebrate neural-specific gains were significantly enriched in categories related to oligodendrocyte differentiation and ensheathment of neurons, consistent with the origin of these glial cells in the gnathostome vertebrate ancestor[40]. In another example, we detected a distinctive enrichment in sensory perception of light stimulus in the octopus skin, probably reflecting the presence of light-activated chromatophore organs all over the cephalopod's body surface[41].

Finally, we focused on the functions of species-specific, tissue-specific gains, which represent 59% of all inferred gains. To identify functional categories potentially over-represented among these species-specific inferences, we plotted a distribution of GOs based on the proportion of their bilaterian-conserved orthogroups experiencing at least one of such species-specific gains (Fig. 5e). Strikingly, the top 5% of this distribution includes cell–cell signalling, tissue development and other developmental categories, which are significantly over-represented in the upper tail compared with the lower percentiles (Fig. 5e and Extended Data Fig. 9c,d). For example, the developmental gene *FGF17* has gained neural specificity during human evolution (Fig. 5f). *FGF17* is a fibroblast growth factor broadly expressed during the embryonic and postnatal brain development of multiple species, but which was co-opted in the adult brain only in humans (Extended Data Fig. 10; data from ref. 18). Remarkably, a recent study[42] showed that the *Fgf17* contained in the cerebrospinal fluid of young mice activates a transcriptional programme leading to proliferation of oligodendrocyte progenitors and, when injected into aged mice, slows down brain aging and improves memory functions (Fig. 5g). Thus, even if *Fgf17*'s potential to induce oligodendrocyte proliferation seems to be ancestral, this gene became part of the adult neural-specific transcriptome only during recent human evolution, where it might contribute to the preservation of cognitive abilities in old age.

Discussion

In this study, we have assembled an extensive dataset of RNA-seq samples spanning 20 bilaterian species and 8 tissues, with the goal of tracing the evolution of gene expression in homologous tissues within and between vertebrates and insects. In terms of phylogenetic range, our study represents a step forward compared with previous works where a similar framework of tissue transcriptional evolution has been applied[17–20], as it extends the investigation range to a large panel of vertebrate and non-vertebrate species, including organisms which diverged ~700 Ma. Indeed, our principal component analysis highlighted consistent transcriptome variation between vertebrate and non-vertebrate species, as perhaps expected given their distinct genomic features (for example, larger genomes and gene numbers, as a consequence of the two WGDs events) and evolutionary traits (for example, longer generation times). This variation manifests into characteristic transcriptomic signatures across tissue types, which are more homogeneous in vertebrates and relatively faster evolving in the other species. Moreover, we designed our study around a symmetric phylogeny for the vertebrate and insect branches. This allowed us to identify not only ancestral features but also parallel, convergent and divergent

https://doi.org/10.1038/s41559-024-02398-5

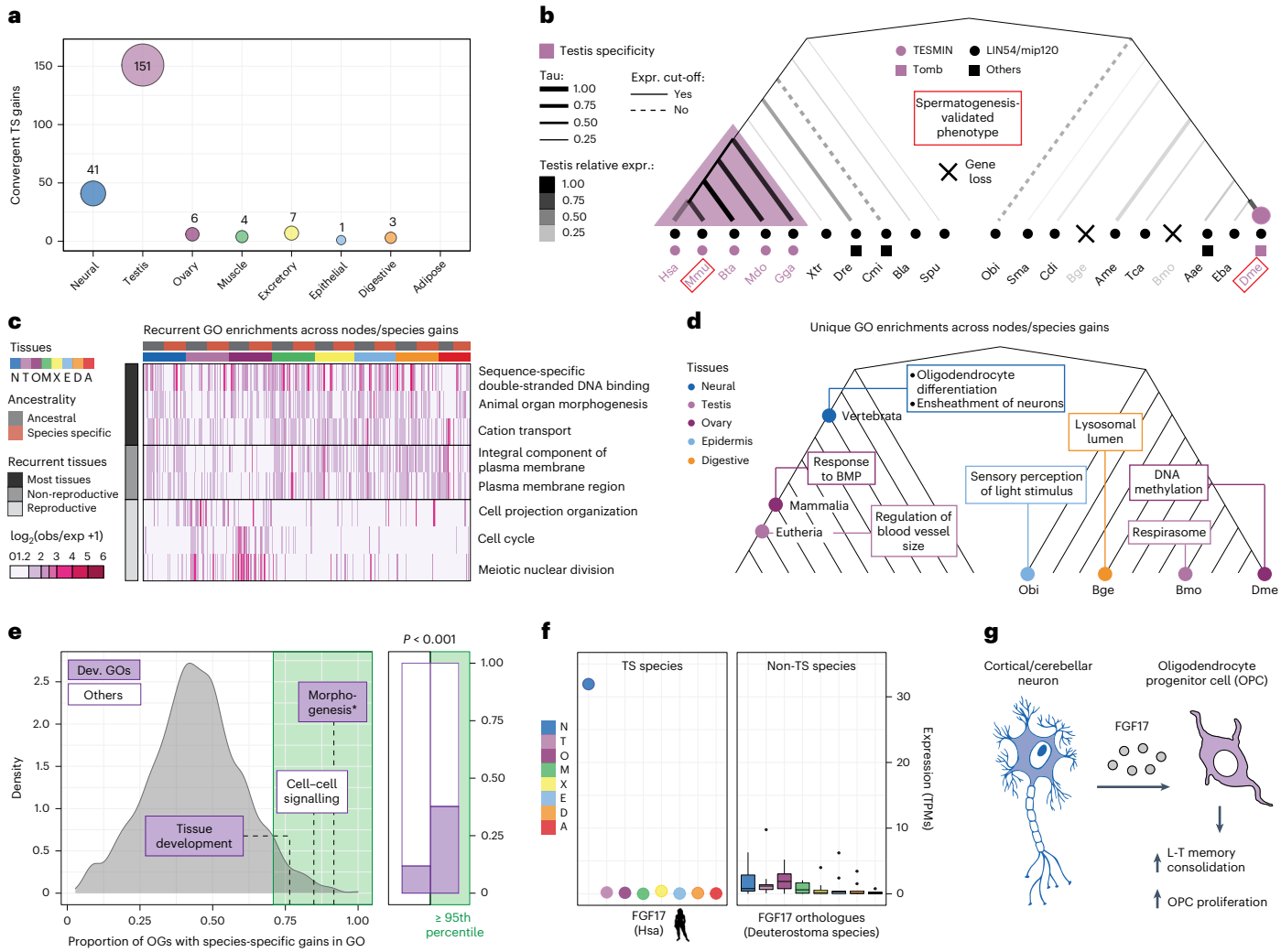


Fig. 5 | Tissue-specific gains are associated with the emergence of unique phenotypes. a, Number of convergent tissue-specificity gains (on the deuterostome and protostome branches) in each tissue. **b**, Example of a convergent testis-specific gain: *TESMIN*/*tomb*. **c**, Heatmap representing GO categories either (1) significantly enriched in the gains of at least 15 nodes/species across all tissues (most/non reproductive labels) or (2) significantly enriched in the gains of at least 8 nodes/species in one tissue exclusively (reproductive label, which indicates ovary and testis combined). The plotted values ($\log_2$(observed/expected + 1)) were computed starting from the proportion of gains in each node/species belonging to the tested category (observed) and the proportion of all bilaterian-conserved orthogroups with a functional annotation belonging to the same category (expected). **d**, Examples of GO categories in each tissue significantly enriched exclusively among the gains of one node/species. **e**, Left: Distribution of the proportion of orthogroups in each GO category with at least one tissue-specific, species-specific gain. The green area represents categories in the 95th percentile or above. Only GO categories including at least 10 bilaterian-conserved orthogroups are shown. Right: Proportions of GO terms below or above the 95th percentile representing developmental functions. The reported *P* value is computed out of the proportions of developmental functions in the 95th percentile coming from 1,000 randomizations of the GO labels (Extended Data Fig. 9c). Morphogenesis* stands for 'anatomical structure formation involved in morphogenesis'. See Methods for definition of developmental categories. **f**, Expression across tissues for human *FGF17* (left; $n = 1$) and its deuterostome orthologues (right; $n = 9$). The centre of the boxplots represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend to the lowest and highest points to a limit of 1.5× the interquartile range from the closest hinge. Outlier points are plotted individually. The human silhouette was downloaded from http://phylopic.org/. **g**, Schematic summary of FGF17's function in the brain (based on ref. 42). BMP, bone morphogenic protein; L-T, long-term; OPC, oligodendrocyte progenitor cell. Neuron icon by Maria Zamchy from thenounproject.com, with modified colours.

evolutionary trajectories of ancestral genes between and within these two major bilaterian lineages. Using this phylogenetic framework, we performed an analysis of the evolutionary dynamics of tissue-specific expression among ancestral bilaterian genes. Strikingly, we found that nearly half of the ancestral bilaterian gene complement has acquired tissue-specific expression in at least one of the studied species, revealing a surprising plasticity for this transcriptomic trait. We thus investigated the timings and mechanisms behind the pervasive evolution of these tissue-specificity patterns, as well as their functional impact.

Before discussing these aspects, however, we acknowledge that a major limitation of our study is the use of bulk RNA-seq data, which merges the signals originating from the different cell types present in each tissue. This is a relevant issue, given that we are analysing distantly related species with divergent tissue histologies. Thus, differences in cell type composition might be a confounding factor in our analyses of gene expression dynamics, especially those involving quantitative comparisons among species across the entire tissue panel (for example, correlations, PCAs and so on). Nevertheless, we aimed at minimizing these considerations by explicitly studying tissue-specific patterns of expression, which are largely qualitative in nature (that is, presence/absence) and thus are more robust to quantitative variations in cell type composition. Indeed, detected changes in tissue-specific transcriptomes

can provide information about evolutionary events such as the origin of novel cell types. For instance, we detected enrichment for oligodendrocyte differentiation exclusively in the neural-specific genes acquired in the vertebrate node, concomitantly with the emergence of this cell type[40].

With regards to evolutionary timings, our phylogenetic inference revealed that most ancestral genes acquired tissue-specific expression during late bilaterian evolution. Despite this, we found that at least ~7% of all ancestral orthogroups have been expressed in a tissue-specific manner since the bilaterian LCA, a number that might be higher if we had investigated more slow-evolving species, such as annelids or other lophotrochozoans. Importantly, all the ancestral tissue-specific modules we identified are linked to core functions within each tissue type, even in those tissues more divergent at the histological and cell type level (for example, digestive system[43]) or mainly originated by convergent morphological trajectories (for example, fat-rich tissues[44]). Neural, muscle and reproductive organ transcriptomes present the largest ancestral tissue-specific modules; this suggests that they have more distinctive and conserved transcriptomic signatures compared with other bilaterian tissues, probably related to the high complexity and specialization of the main cell types that form them (neurons, myocytes and meiotic cells, respectively).

At the mechanistic level, we showed that tissue-specific gains have a strong association with gene duplication across the entire bilaterian phylogeny, as previously reported for more limited lineages[37,45]. Furthermore, we investigated how often evolution of tissue specificity occurred through specialization, by which the specialized paralogue in a given species reduced its expression in the tissues without tissue specificity compared with the broadly expressed ancestral patterns, which are preserved in the non-tissue-specific paralogues of that species as well as the orthologues in the other species. This pattern had previously been identified for more restricted groups, including paralogues originated from vertebrate[37] and salmon[46] WGDs, together with gene duplicates specific in the pea aphid[47], primates and rodents[48]; here we expanded the search space and provided evidence that specialization is associated with tissue-specificity gains throughout the entire bilaterian phylogeny. Another remarkable finding in the context of gene duplication is the seemingly prolonged effect of the vertebrate WGDs on the amount of tissue-specificity gains throughout recent vertebrate evolution. Specifically, the Vertebrata node showed the highest level of tissue-specificity gains in the phylogeny, particularly among paralogues retained from the WGDs (ohnologues), as expected from a direct causal effect of these events. However, subsequent ancestral nodes and extant species within the vertebrate lineage also exhibited higher number of gains, as well as a higher fraction of affected ohnologues, compared with other phylogenetically equivalent non-vertebrate nodes and species. While this could partly be due to loss of tissue specificity in early branching vertebrate species, we suggest that this pattern reflects the increased likelihood of orthogroups with retained ohnologues in vertebrates to evolve tissue specificity even millions of years after the WGDs that generated the genetic redundancy. If so, this unexpected finding implies that the evolutionary impact of WGDs on phenotypic diversification may go beyond immediate subsequent effects, providing an additional potential explanation for the lag observed between the timing of WGDs and their purported consequences in multiple lineages[49–51].

Finally, we assessed the functional impact of the rewiring of tissue-specific transcriptomes. Almost 60% of our inferred tissue-specificity gains occurred in specific species and were often associated with the emergence of unique phenotypes, highlighting the potential of novel tissue-specific expression patterns to underlie organismal novelties[52]. For instance, we detected a distinctive enrichment in sensory perception of light stimulus in the octopus skin, consistent with the unique presence of light-activated chromatophore organs all over the cephalopod's body surface[41]. We also uncovered a significant tendency for developmental genes to retain adult tissue-specific expression in a species-specific manner. Cases such as *FGF17* in humans that we reported here point to a potential widespread, functional co-option of ancestral developmental genes within distinct tissue-specific transcriptomes throughout the most recent bilaterian evolution. Future research should elucidate the functional relevance of the adult tissue-specific expression of these developmental genes as well as of the myriad of other tissue-specific genes we identified and how they ultimately contribute to animal evolution.

Methods

Gene orthology calls

We used Broccoli (v.1.2)[53] to infer gene orthogroups among all protein-coding genes from the 20 species. To avoid redundant gene homology calls, we selected one representative protein isoform for each gene in each species (that is, the isoform with the longest coding sequence). See Supplementary Methods for details on the genome annotation and sequence files, Extended Data Fig. 1 and Supplementary Data 2 for gene orthogroup statistics, and the Supplementary Dataset for gene orthogroup files.

RNA-seq sample dataset

We downloaded a total of 1,136 individual RNA-seq samples across 18 species. All downloaded samples and relative information can be found in Supplementary Data 3. Moreover, we generated 89 RNA-seq samples for 15 species covering different tissues that were missing in public resources. See Supplementary Data 3 for more details (all the samples generated for this project report 'in_house' in the Sample_origin field). All these samples were dissected from adult animals and the RNA extracted using the most suited protocol for the organism and tissue, namely TRIzol reagent (Thermo Fisher) or Qiagen RNeasy kit (QIAGEN) (see GEO series GSE205498 for more details on sample extraction and processing protocols). These RNA samples were used to construct standard Illumina RNA-seq libraries at the CRG Genomics Unit, and an average of ~78 million 125-nucleotide paired-end reads were generated for each of them in a HiSeq2500 sequencing system. In the case of octopus, the sequencing was performed at the University of Chicago with a NovaSeq system. In total, we generated ~7.6 billion individual reads. FastQC reports for all in-house-generated samples are available in the Supplementary Dataset. Extra metadata information for both in-house-generated and publicly available samples is provided in Supplementary Data 3.

RNA-seq quantification

We quantified expression using Kallisto quant[54], setting the parameters '-b 100–single -l 190 -s 20' for single-end RNA-seq samples and '-b 100' for paired-end RNA-seq samples (with -b = number of bootstrap samples; -l = estimated average fragment length; -s = estimated standard deviation of fragment length). Kallisto logs for all samples are provided in the Supplementary Dataset. For each species, we quantified gene expression for each sample by summing the raw counts of all its corresponding annotated transcripts. We next normalized the expression with DESeq2 (ref. 55) and used the effective lengths returned by Kallisto to compute the transcripts per million (TPMs). For all analyses, $\log_2(\mathrm{TPM}+1)$ was used as the final expression measure for each sample.

Metasamples and tissue expression measures

When multiple datasets were available for a given tissue and species, we grouped the RNA-seq samples into a maximum of three metasamples. This was done to: (1) increase read depth per metasample, (2) dilute potential batch effects from publicly available samples and (3) facilitate downstream analyses and comparisons by having a comparable number of replicates across tissues and species. Metasample groups are detailed in the column 'Metasample' in Supplementary Data 3. In particular, we followed the approach that we previously described for

human, mouse, cow, zebrafish and fruit fly in VastDB (https://vastdb.crg.eu/)[56], where samples from comparable experiments based on clustering approaches are pooled. For this study, we first computed the median expression across all the samples included in each metasample, which we used as its representative measure. Then, we calculated the average of these measures across all the metasamples belonging to each tissue, generating representative expression values at the tissue level. The same expression quantification procedure was adopted for both in-house-generated and publicly available samples. Importantly, the use of alternative combinations of RNA-seq samples to define metasamples, computed either after averaging the expression of all available samples for each tissue or after randomizing the samples of each tissue across the relative metasamples, yielded similar results (see Supplementary Methods).

PCA and clustering analysis

To investigate the interrelation among our metasamples, we performed a PCA on the best-ancestral orthogroups normalized expression matrix (see Supplementary Methods) using the prcomp function in R (centre = TRUE, scale = TRUE). To assess the biological nature of each principal component, we performed one-sided analysis of variance (ANOVA) tests between species and tissue groups, employing the aov function in R (Extended Data Fig. 3c; shown *P* values were Bonferroni corrected). The heatmap in Extended Data Fig. 3d was generated using the pheatmap R package with the ward.D2 clustering method on the best-ancestral orthogroups *z*-scored expression matrix (see Supplementary Methods).

Definition of ancestral bilaterian tissue-specific modules

As summarized in Fig. 2a, we first performed an sPLS-DA with the splsda function in the mixOmics package[57] in R, using as input the best-ancestral orthogroups normalized expression matrix (but where all bilaterian-conserved orthogroups are included). We specifically compared all tissue groups versus each other, selecting the optimal number of components and loadings per component by running the tune.splsda function on the same expression table with the following parameters: ncomp = 10, validation = 'Mfold', folds = 4, dist = 'max.dist', measure = 'BER', test.keepX = c(1:10, seq(20, 300, 10)), nrepeat = 10. Since each of the resulting components specifically separated the metasamples of each tissue group (Extended Data Fig. 4a–f), we used the corresponding loadings (which represent orthogroups with the most distinctive expression profiles in the isolated tissue compared with the others) to define the respective ancestral bilaterian tissue-specific modules. Importantly, contrary to a PCA, the proportion of variance explained by consecutive components does not necessarily decrease, as the aim is not to maximize the variance. As an extra filter, we further selected only those best-ancestral orthogroups that had the highest median expression in the isolated tissue both among vertebrates and insects. To be able to pool tissue expression values across species, we considered the *z*-scored expression matrix described in Supplementary Methods (see 'Best-ancestral orthogroups normalized and *z*-scored expression matrices'), but where all bilaterian-ancestral orthogroups are included. The values plotted in Fig. 2c,d and Extended Data Fig. 4g–l correspond to the median of these expression measures among all vertebrates, all insects or all outgroups (that is, only 3 values instead of 20 are plotted per orthogroup and tissue).

Analysis of ancestral bilaterian tissue-specific modules

GO enrichment analyses were performed with the gprofiler2 (ref. 58) R package, using either the human or the fruit fly ontology transfers as GO annotation and all bilaterian-conserved, best-ancestral orthogroups as background. All *P* values were false discovery rate (FDR) corrected. Results obtained with both GO annotations are provided in Supplementary Data 6 (human) and 7 (fruit fly), but only GO enrichments from the human annotations are discussed in the relative Results section and represented in Fig. 2e,f,i. For the representation of GO networks of significantly enriched categories (adjusted $P \leq 0.05$) in Fig. 2i, only significant categories containing at least 5 genes in the tested set were considered. The networks were obtained from Revigo (http://revigo.irb.hr/)[59], selecting large output lists (90% of the input list; option 0.9) for all modules except the neural-differential (for which 0.4 [40%] was selected). To characterize the phenotypic impact of these genes, we downloaded all validated gene–phenotype associations from Ensembl[60] (v.105) for human and mouse and from FlyBase[61] as updated in January 2020. Neural phenotypes were defined as anything matching 'neuro', 'behaviour', 'brain', 'glia' or 'CNS' (case insensitive), while testis phenotypes (which we reasoned should also include broader reproduction-related phenotypes) were defined as anything matching 'sperm', 'infert', 'sterile' or 'testis' (case insensitive). Orthogroups with positive matches in either species were considered for the plots shown in Fig. 2g,h. In this analysis, no distinctions were made between genes lacking a neural/testis phenotype and genes without phenotypic characterization. Neural and testis phenotypes associated with the respective ancestral tissue-specific module are reported in Supplementary Data 5, while all phenotypic associations mapped to the respective bilaterian-conserved orthogroup are available in the Supplementary Dataset. Finally, the PCAs in Supplementary Fig. 4a,b were performed on the best-ancestral orthogroups normalized expression matrix (see Supplementary Methods) but after filtering either for all the orthogroups belonging to the ancestral bilaterian tissue-specific modules (Supplementary Fig. 4a) or for a matched number of orthogroups from each tissue's modules (maximum of 20 random orthogroups per module; Supplementary Fig. 4b).

Tissue-specificity calls

To perform the tissue-specificity calls, we first computed the Tau[35] for all genes separately in each species. Tau is a measure of tissue specificity ranging from 0 (ubiquitous genes) to 1 (highly tissue-specific genes). For each species, we employed as input a quantile-normalized expression matrix of $\log_2$(TPMs + 1) values averaged by tissue (that is, one value per tissue). We defined as tissue specific in each species all genes with Tau ≥ 0.75 and maximum expression ≥$\log_2(5)$. The Tau threshold was chosen by looking at the general Tau distributions across species (Fig. 3a and Supplementary Fig. 5), many of which show a bimodal trend where a Tau cut-off of 0.75 would select the majority of the upper tail (that is, highly tissue-specific genes); moreover, this threshold is similar to implemented tissue-specificity thresholds in previous publications[62–64]. To associate these tissue-specific genes with one or two tissues ('Associated tissue(s)' in Fig. 3b, Supplementary Fig. 6 and Extended Data Fig. 2c), we evaluated the expression proportion per tissue (tissue_expr/all_tissue_expr), where 'tissue_expr' is the average normalized $\log_2$(TPMs + 1) expression of the gene in the target tissue and 'all_tissue_expr' is the sum of the average normalized $\log_2$(TPMs + 1) expression values across all tissues. Specifically, we applied the following steps for each gene in each species (Extended Data Fig. 2c): (1) if the difference in expression proportion between the two most highly expressed tissues was ≥0.10 and their ratio ≥1.7, we associated the gene only with the top tissue. (2) If the above conditions were not fulfilled, but the difference in expression proportion between the second and third most highly expressed tissues was ≥0.15, we associated the gene with the two top tissues (double tissue specificity). (3) Otherwise, the gene was not considered as tissue specific and not associated with any tissue. In addition, for the gain/loss inferences (see next section), we more loosely defined the 'Top tissue(s)', corresponding to the 'Associated tissue(s)', when available, or simply to the two tissues with the highest expression (Extended Data Fig. 2c, last example).

Phylogenetic inference of tissue-specificity gains

We performed the phylogenetic inferences of tissue-specificity gains and losses for each tissue separately, considering all the

orthogroups presenting at least one tissue-specific call in that tissue (see 'tissue-specificity calls'). We implemented two subsequent ad hoc, parsimony-based inference approaches independently for each major branch (deuterostome and protostome), which we developed due to the limitations of other inference methods with respect to our dataset and our specific scientific aim (see Supplementary Discussion). First, we applied a 'strict approach', inferring a maximum of one tissue-specificity gain for each major branch. Here we inferred a gain in a node if (Extended Data Fig. 5a left): (1) the first-branching species in the node was tissue specific in the query tissue (as defined in the previous section); (2) at least 50% of the node's descendant species with an orthologue had Tau ≥ 0.60 and were associated with the query tissue; and (3) none of the outgroup species to that node on the same branch that passed the expression cut-off had Tau ≥ 0.60 and were associated with the query tissue. Exceptionally, in the case of the most internal nodes (that is, Euarchontoglires: human and mouse; Cyclorrhapha: fruit fly and hoverfly), we required Tau ≥ 0.6 and association with the query tissue in both species, and a tissue-specific call in that tissue for at least one of them.

Second, for all the orthogroups that could not be classified with the first strict approach for a given branch, we inferred gains with less stringent requirements ('relaxed approach'; Extended Data Fig. 5a right). Here we inferred gains in the last common ancestor of all species with Tau ≥ 0.60 that are associated with the query tissue as long as at least one tissue-specific gene is present. However, the relaxed approach inferred multiple gains on each branch if the minimum distance between two species or nodes respecting those tissue-specificity cut-offs was higher than 3 nodes (for example, in human and in chicken, or in Eutheria and in zebrafish). Also, if no inference of gain in an ancestral node could be done by either approach, tissue-specific genes (as defined in the previous section) were considered species-specific gains. Finally, from the combined output of both approaches, we inferred an ancestral bilaterian (or earlier) tissue specificity whenever a 'gain' was identified in both Deuterostoma/Chordata/Vertebrata and Protostoma/Arthropoda/Insecta with either strict or relaxed criteria (Extended Data Fig. 5b; 'merged' label in Supplementary Data 9). As an exception, since shark testis samples showed poor correlation with other testis samples, we also inferred ancestral bilaterian tissue specificity for testis in case of gain inferences in Euteleostomi and Protostoma.

Phylogenetic inference of tissue-specificity losses

We then inferred tissue-specificity losses exclusively starting from the nodes in which gains were inferred for each tissue (Extended Data Fig. 5c). In case of ancestral bilaterian tissue specificity, the inferences were conducted separately on the deuterostome and protostome branches. We considered as potential losses all species (internal to the node with the inferred gain) where either: (1) Tau ≤ 0.45; (2) the query tissue was not among the top tissue(s), as defined above (Extended Data Fig. 2c); or (3) the difference in expression proportions between the query tissue and the third highest tissue was ≤0.1. Then, starting from the innermost species with a potential loss, if there were two or more consecutive such species, we inferred a loss in the node corresponding to their LCA and a novel gain in the node of the LCA of their consecutive inner species if: (1) all these species were tissue specific as described above, (2) the ancestral loss was separated by at least one node from the most ancestral gain and (3) the total number of these new inferences (including single losses in all the species excluded from the ancestral loss inference) was lower than the number of original inferences (that is, independent losses for each potential loss species). Otherwise, separated losses for each single species were inferred.

Duplication and specialization of tissue-specificity gains

Each orthogroup's duplicated proportion was defined as the number of species with at least two paralogues over the total number of considered species (which depends on the tested node). The mean duplicated proportion for the orthogroups with gains in each node compared to the relative background (that is, all orthogroups in that node) is shown in Fig. 4d. The proportion of orthogroups with gains including 2R-ohnologues (Fig. 4c) was based on the list of 2R-ohnologues provided in ref. 65. The ten randomized bilaterian-conserved gene orthogroups used in Extended Data Fig. 8a were obtained by shuffling genes within each species while preserving the original paralogy structures (that is, each randomized orthogroup conserved the original number of paralogues from each species, but the actual orthologous genes no longer corresponded across species).

We then checked how each tissue-specific gain fitted the specialization hypothesis. We started from the same expression matrices used for the tissue-specificity call (see above), comparing the median expression in each tissue between species with tissue specificity and species without tissue specificity (including species with inferred tissue-specificity losses). For each gain, we counted for how many tissues (excluding the tissue with tissue specificity) this median expression was higher in the species without tissue specificity (specialization-supporting tissues, ranging 0–7; relative proportions across nodes and species in Fig. 4f and Extended Data Fig. 7d,e). For the gains in each node and species, we performed 100 randomizations of the tissue-specificity labels among all species in the relative orthogroup. For each of these randomization rounds, we counted the proportion of gains in which the number of specialization-supporting tissues was ≥5. We plotted in purple the distributions of these proportions for all randomizations, overlaying the relative observed distributions in Extended Data Fig. 7d,e or their collapsed distributions in Fig. 4f.

Functional characterization of tissue-specificity gains

Parallel and convergent gains of tissue specificity (Extended Data Fig. 9a and Fig. 5a) were evaluated exclusively among those best-TS orthogroups that present tissue-specificity gains in only one tissue on each of the main branches (deuterostome or protostome). The GO enrichment analysis on the orthogroups with gains in each node/species reported in Fig. 5c,d and Supplementary Data 10 and 13 were performed as described in 'Analysis of ancestral bilaterian tissue-specific modules' and using the GO transfers derived from the human annotation. The same enrichments were also repeated using the vertebrate-specific and insect-specific GO transfers (Supplementary Data 11 and 12). For the heatmap in Fig. 5c, we exclusively considered GO categories that were either (1) significantly enriched in the gains of at least 15 nodes/species across all tissues or (2) significantly enriched in the gains of at least 8 nodes/species in one tissue exclusively; in this last analysis, ovary and testis were grouped to catch a combined signature from the reproductive organs. The plotted values ($\log_2$(observed/expected + 1)) were computed starting from the proportion of gains in each node/species belonging to the tested category (observed) and the proportion of all bilaterian-conserved orthogroups with a functional annotation belonging to the same category (expected). Highly redundant categories were manually removed. For Fig. 5d and Supplementary Data 13, we only considered the GO categories that were exclusively enriched in one node or species. Then, we moved to the characterization of species-specific gains, where we evaluated whether developmental GOs were more represented in these recent gains compared with ancestral ones. Developmental GO categories were defined starting from the human transferred GO annotation (see Supplementary Methods) as any term matching 'develop', 'differentiation', 'determination', 'morphogen', 'commitment', 'specification', 'regionalization', 'formation' or 'genesis'. For the plot shown in Fig. 5e, only the GO categories including at least 10 bilaterian-conserved orthogroups were considered. The gene set enrichment analysis in Extended Data Fig. 9c was performed with the fgsea package in R[66], and the distribution shown in Extended Data Fig. 9d resulted from 1,000 randomizations of the GO categories labels across the proportions of orthogroups in each category that included at least one species-specific gain.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The FASTQ and processed files of the RNA-seq samples generated for this project are available at GEO under series GSE205498. The Supplementary Dataset is available via Mendeley Data at https://doi.org/10.17632/22m3dwhzk6.2 (ref. 67).

Code availability
All code used for analysis and figure generation is available on GitHub at https://github.com/fedemantica/bilaterian_GE (ref. 68).

References

1. Evans, S. D., Hughes, I. V., Gehling, J. G. & Droser, M. L. Discovery of the oldest bilaterian from the Ediacaran of South Australia. *Proc. Natl Acad. Sci. USA* **117**, 7845–7850 (2020).
2. Brusca, R. C., Moore, W. & Shuster, S. M. *Invertebrates* 345–372 (Sinauer Associates, 2016).
3. Paps, J. & Holland, P. W. H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat. Commun.* **9**, 1730 (2018).
4. Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nat. Ecol. Evol.* **4**, 524–533 (2020).
5. Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of *Homo sapiens*. *Genome Biol.* **9**, R33 (2008).
6. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
7. King, M.-C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
8. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
9. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18**, 53–80 (2002).
10. Arntfield, M. E. & van der Kooy, D. β-Cell evolution: how the pancreas borrowed from the brain: the shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. *Bioessays* **33**, 582–587 (2011).
11. Almudi, I. et al. Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. *Nat. Commun.* **11**, 2631 (2020).
12. Clark-Hachtel, C. M. & Tomoyasu, Y. Two sets of candidate crustacean wing homologues and their implication for the origin of insect wings. *Nat. Ecol. Evol.* **4**, 1694–1702 (2020).
13. Bruce, H. S. & Patel, N. H. Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. *Nat. Ecol. Evol.* **4**, 1703–1712 (2020).
14. Martín-Durán, J. M. et al. Convergent evolution of bilaterian nerve cords. *Nature* **553**, 45–50 (2018).
15. Thomas, J. A., Welch, J. J., Lanfear, R. & Bromham, L. A generation time effect on the rate of molecular evolution in invertebrates. *Mol. Biol. Evol.* **27**, 1173–1180 (2010).
16. Wyder, S., Kriventseva, E. V., Schröder, R., Kadowaki, T. & Zdobnov, E. M. Quantification of ortholog losses in insects and vertebrates. *Genome Biol.* **8**, R242 (2007).
17. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
18. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
19. Chen, J. et al. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.* **29**, 53–63 (2019).
20. Fukushima, K. & Pollock, D. D. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nat. Commun.* **11**, 4459 (2020).
21. Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
22. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).
23. Burkhardt, P. & Sprecher, S. G. Evolutionary origin of synapses and neurons – bridging the gap. *Bioessays* **39**, 1700024 (2017).
24. Sebé-Pedrós, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534.e20 (2018).
25. Inaba, K. Sperm flagella: comparative and phylogenetic perspectives of protein components. *Mol. Hum. Reprod.* **17**, 524–538 (2011).
26. Daldello, E. M., Luong, X. G., Yang, C.-R., Kuhn, J. & Conti, M. Cyclin B2 is required for progression through meiosis in mouse oocytes. *Development* **146**, dev172734 (2019).
27. Li, J., Ouyang, Y.-C., Zhang, C.-H., Qian, W.-P. & Sun, Q.-Y. The cyclin B2/CDK1 complex inhibits separase activity in mouse oocyte meiosis I. *Development* **146**, 648053 (2019).
28. Zeng, Y. et al. Bi-allelic mutations in MOS cause female infertility characterized by preimplantation embryonic arrest. *Hum. Reprod.* **37**, 612–620 (2022).
29. Tay, J., Hodgman, R., Sarkissian, M. & Richter, J. D. Regulated CPEB phosphorylation during meiotic progression suggests a mechanism for temporal control of maternal mRNA translation. *Genes Dev.* **17**, 1457–1462 (2003).
30. Gąsiorowski, L. et al. Molecular evidence for a single origin of ultrafiltration-based excretory organs. *Curr. Biol.* **31**, 3629–3638.e2 (2021).
31. Thakurela, S. et al. Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov.* **2**, 15045 (2016).
32. Eckler, M. J. & Chen, B. Fez family transcription factors: controlling neurogenesis and cell fate in the developing mammalian nervous system. *Bioessays* **36**, 788–797 (2014).
33. Taylor, M. V. & Hughes, S. M. Mef2 and the skeletal muscle differentiation program. *Semin. Cell Dev. Biol.* **72**, 33–44 (2017).
34. Mathiyalagan, N. et al. Meta-analysis of grainyhead-like dependent transcriptional networks: a roadmap for identifying novel conserved genetic pathways. *Genes* **10**, 876 (2019).
35. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
36. Roelofs, D. et al. Multi-faceted analysis provides little evidence for recurrent whole-genome duplications during hexapod evolution. *BMC Biol.* **18**, 57 (2020).
37. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
38. Oji, A. et al. Tesmin, metallothionein-like 5, is required for spermatogenesis in mice. *Biol. Reprod.* **102**, 975–983 (2020).
39. Jiang, J., Benson, E., Bausek, N., Doggett, K. & White-Cooper, H. Tombola, a tesmin/TSO1-family protein, regulates transcriptional activation in the *Drosophila* male germline and physically interacts with always early. *Development* **134**, 1549–1559 (2007).
40. Hines, J. H. Evolutionary origins of the oligodendrocyte cell type and adaptive myelination. *Front. Neurosci.* **15**, 757360 (2021).
41. Ramirez, M. D. & Oakley, T. H. Eye-independent, light-activated chromatophore expansion (LACE) and expression of phototransduction genes in the skin of *Octopus bimaculoides*. *J. Exp. Biol.* **218**, 1513–1520 (2015).

42. Iram, T. et al. Young CSF restores oligodendrogenesis and memory in aged mice via Fgf17. *Nature* **605**, 509–515 (2022).
43. Hartenstein, V. & Martinez, P. Structure, development and evolution of the digestive system. *Cell Tissue Res.* **377**, 289–292 (2019).
44. Ottaviani, E., Malagoli, D. & Franceschi, C. The evolution of the adipose tissue: a neglected enigma. *Gen. Comp. Endocrinol.* **174**, 1–4 (2011).
45. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. *PLoS Comput. Biol.* **12**, e1005274 (2016).
46. Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
47. Fernández, R. et al. Selection following gene duplication shapes recent genome evolution in the pea aphid *Acyrthosiphon pisum*. *Mol. Biol. Evol.* **37**, 2601–2615 (2020).
48. Farré, D. & Albà, M. M. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol. Biol. Evol.* **27**, 325–335 (2010).
49. Clark, J. W. & Donoghue, P. C. J. Constraining the timing of whole genome duplication in plant evolutionary history.*Proc. Biol. Sci.* **284**, 20170912 (2017).
50. Macqueen, D. J. & Johnston, I. A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. Biol. Sci.* **281**, 20132881 (2014).
51. Donoghue, P. C. J. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20**, 312–319 (2005).
52. Almudí, I. & Pascual-Anaya, J. in *Old Questions and Young Approaches to Animal Evolution* (eds Martín-Durán, J. M. & Vellutini, B. C.) 107–132 (Springer, 2019).
53. Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
54. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
55. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
56. Tapial, J. et al. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* **27**, 1759–1768 (2017).
57. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
58. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2—an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.*F1000Research* **9**, ELIXIR-709 (2020).
59. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
60. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
61. Gramates, L. S. et al. FlyBase: a guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).
62. Jin, L. et al. A pig BodyMap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription. *Nat. Commun.* **12**, 3715 (2021).
63. Wang, Z.-Y. et al. Transcriptome and translatome co-evolution in mammals. *Nature* **588**, 642–647 (2020).
64. Guschanski, K., Warnefors, M. & Kaessmann, H. The evolution of duplicate gene expression in mammalian organs. *Genome Res.* **27**, 1461–1474 (2017).
65. Touceda-Suárez, M. et al. Ancient genomic regulatory blocks are a source for regulatory gene deserts in vertebrates after whole-genome duplications. *Mol. Biol. Evol.* **37**, 2857–2864 (2020).
66. Korotkevich, G. et al. Fast gene set enrichment analysis. Preprint at *bioRxiv* https://doi.org/10.1101/060012 (2021).
67. Mantica, F. & Irimia, M. Pervasive evolution of tissue-specificity of ancestral genes differentially shaped vertebrates and insects, V2. *Mendeley Data* https://doi.org/10.17632/22m3dwhzk6.2 (2023).
68. fedemantica. bilaterian_GE. *GitHub* https://github.com/fedemantica/bilaterian_GE (2023).
69. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

Acknowledgements

We thank Q. T. Ramon for the original drawing of tissue icons; N. Arecco, N. B. Morais, A. Sebé-Pedrós and D. Weghorn for critical feedback on the manuscript; and the CRG Genomics Unit for the RNA sequencing. This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC-StG-LS2-637591 and ERCCoG-LS2-101002275 to M.I.), by the Spanish Ministry of Economy and Competitiveness (BFU-2017-89201-P and PID2020-115040GB-I00 to M.I.) and by the 'Centro de Excelencia Severo Ochoa 2013-2017'(SEV-2012-0208). F.M. holds a FPI fellowship associated with the grant BFU-2017-89201-P. Additional support for this research was provided by the Spanish MINECO (PGC2018-098427- B-I00 to D.M. and X.F.-M.), the Czech Science Foundation (22-21244S to M.N.), the Australian Research Council (grant DP200103219 to P.D.C. and F.T.) and the National Institutes of Health-NIAID (grant R21AI167849 to F.G.N.).

Author contributions

F.M. performed most analyses and generated most figures and tables. L.P.I. built the motif dataset, designed and performed all motif-related analysis, and contributed to intellectual discussion. Y.M. and A.T.-M. performed additional analyses and contributed to intellectual discussion. J.P., A.T.-M., J.C., X.F.-M., F.T., D.B., S.B., T.D., M.N., P.D.C., F.G.N., H.E., M.I.A., C.B.A., K.R.W., I.A. and D.M. contributed RNA and/or tissue samples. F.M. and M.I. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

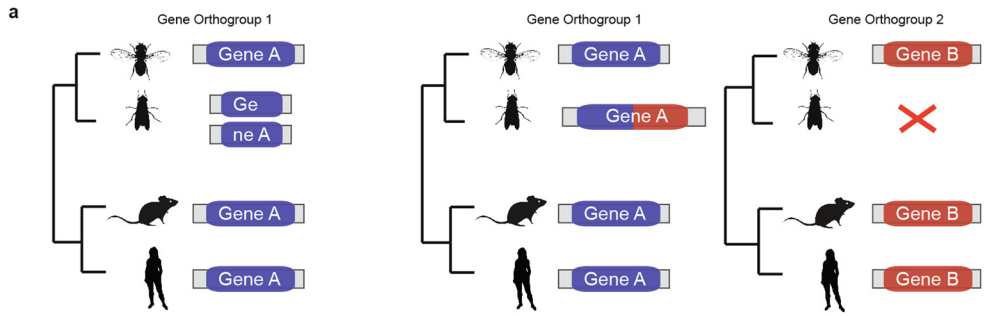
Extended data is available for this paper at https://doi.org/10.1038/s41559-024-02398-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-024-02398-5.

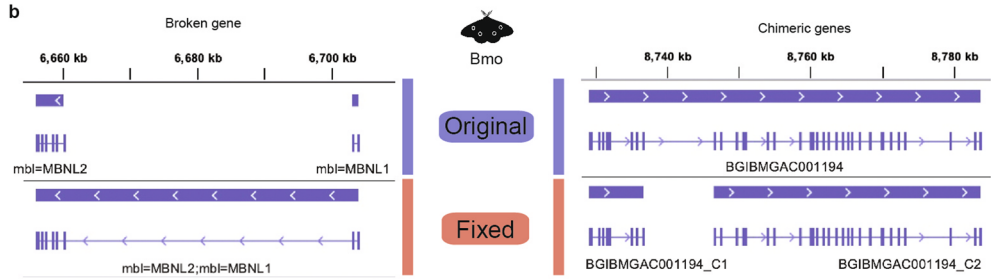
Correspondence and requests for materials should be addressed to Manuel Irimia.

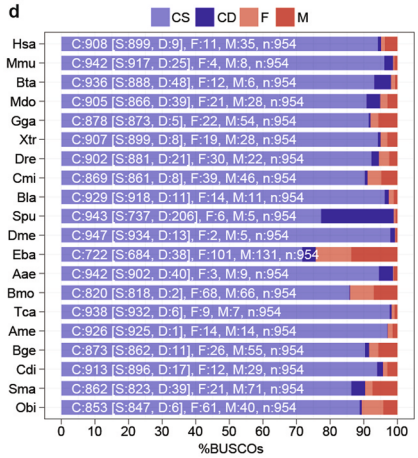
Peer review information *Nature Ecology & Evolution* thanks Marie Sémon, Emily Wong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

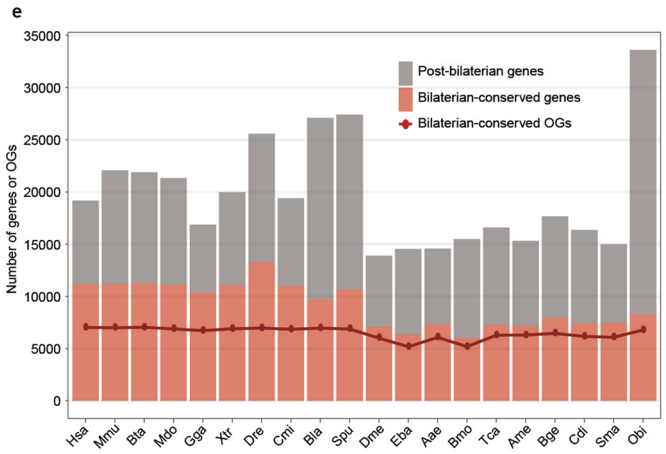
https://doi.org/10.1038/s41559-024-02398-5

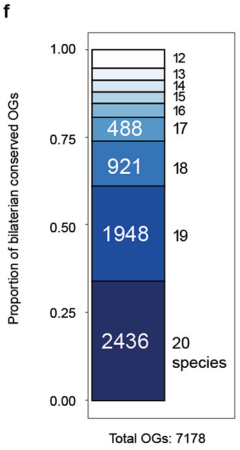


Species	Resolved broken groups	Resolved chimeric genes	Unresolved chimeric genes
Bta	132	4	2
Mdo	147	21	27
Gga	246	5	11
Xtr	88	5	24
Dre	103	0	5
Cmi	335	5	9
Bla	77	8	24
Spu	365	1	8
Eba	95	157	220
Aae	38	2	5
Bmo	329	16	40
Tca	72	57	56
Ame	184	11	26
Bge	120	11	18
Cdi	54	133	96
Sma	204	160	139
Obi	478	0	1



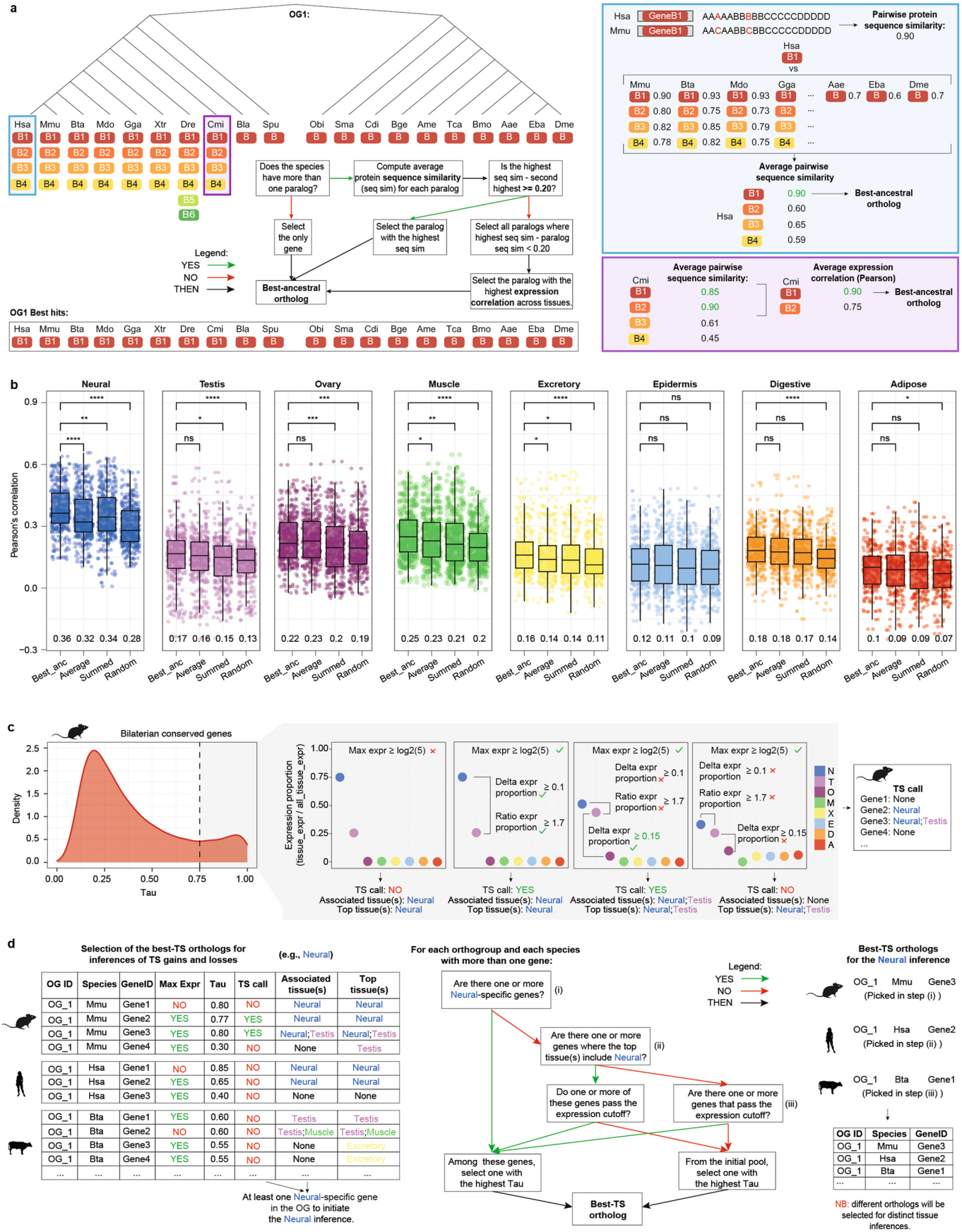






Extended Data Fig. 1 | Gene annotation refinements and statistics of bilaterian-conserved orthogroups. a. Schematic representation of broken (left) and chimeric (right) genes and how they potentially influence gene orthology inferences. Animal silhouettes were downloaded from http://phylopic.org/. Credits to Gareth Monger for the hoverfly icon (https://creativecommons.org/licenses/by/3.0/). **b**. Examples of a broken (left) and chimeric (right) genes corrected in the silkworm gene annotation. **c**. Statistics of corrected and unresolved broken and chimeric genes across all species. **d**. Results from a *BUSCO* run (options *-m proteins -l metazoa_odb10*) assessing the status of 954 metazoa single-copy orthologs in the proteomes of all the species. CS: complete and single-copy, CD: complete and duplicated, F: fragmented, M: missing. **e**. Barplot representing the number of bilaterian-conserved (red) or more recent (grey) protein-coding genes across all species. The line plot represents the number of bilaterian-conserved orthogroups (OGs; that is, orthogroups conserved in at least 12 species) in which genes from each species are represented. **f**. Proportions of bilaterian-conserved orthogroups based on the number of species in which they are conserved.

https://doi.org/10.1038/s41559-024-02398-5

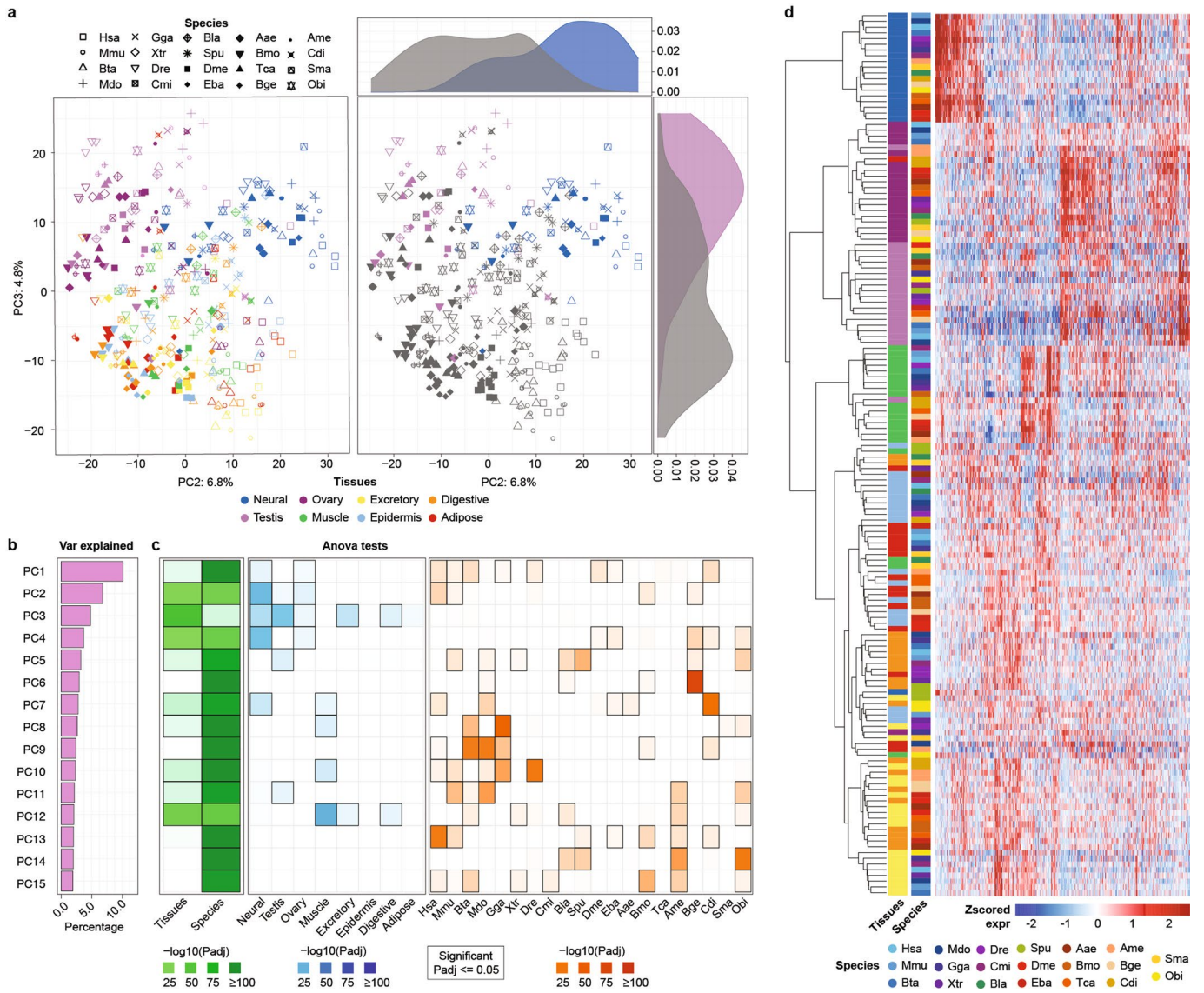


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41559-024-02398-5

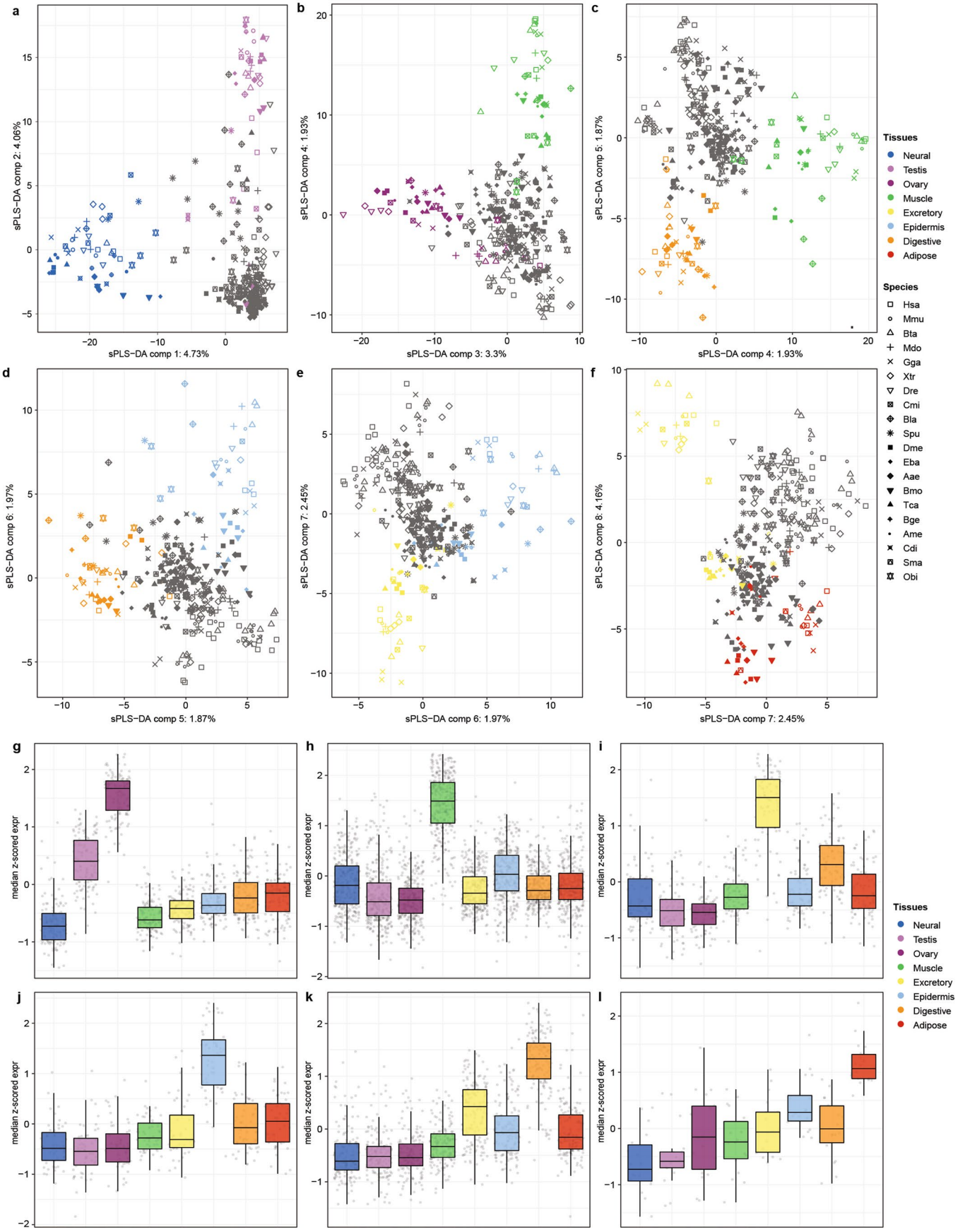
Extended Data Fig. 2 | Definition of best-ancestral and best-TS orthogroups. **a**. Schematic and relative example for the selection of bilaterian-conserved, best-ancestral orthologs in each species and tissue (see Supplementary Methods). **b**. Distributions of Pearson's correlation coefficients from all intra-tissue, species pairwise comparisons of gene expression upon distinct procedures for paralog selection and gene expression quantification. The expression measure for each species in each orthogroup corresponds to the expression of its best-ancestral ortholog (Best_anc), the average expression among all its paralogs (Average), the summed expression among all its paralogs (Summed) and the expression of a randomly selected paralog (Random). Significance levels of two-sided Wilcoxon rank-sum tests comparing the Best_anc distribution to each of the others are reported at the top, while the median value of each distribution is printed at the bottom. Correlations are performed on z-scored expression matrices (see Supplementary Methods). Only the 2,436 gene orthogroups conserved in all species (n = 20) were considered. P-value significance levels are defined as follows: **** = p-value ≤ 0.0001, *** = p-value ≤ 0.001, ** = p-value ≤ 0.01, * = p-value ≤ 0.05. The boxplot features are defined as follows: the center line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend respectively to the lowest and highest points, to a limit of 1.5 multiplied by the interquartile range from the closest hinge. c. Schematic of the procedure adopted to associate all tissue-specific genes in each species (Tau ≥ 0.75) with the tissue(s) with tissue-specificity. This association (which we also evaluated for non-tissue-specific genes) will be considered for the inference of tissue-specificity gains (Extended Data Fig. 5). Additionally, we identified the top tissue(s) (that is, the tissue(s) with the highest expression) for all bilaterian-conserved genes, which will be considered for the selection of the best-TS orthogroups and the inference of tissue-specificity losses in each tissue (panel d and Extended Data Fig. 5c, respectively). **d**. Schematic and relative example for the selection of the best-TS ortholog in each species (see Supplementary Methods). Animal silhouettes were downloaded from http://phylopic.org/.

https://doi.org/10.1038/s41559-024-02398-5



Extended Data Fig. 3 | Partial conservation of tissue-specific expression profiles among ancestral bilaterian genes. a. Coordinates of the second (PC2; x axis) and third (PC3; y axis) components of a PCA performed on the best-ancestral orthogroups normalized gene expression matrix. Only the 2,436 best-ancestral orthogroups conserved in all species were considered. Tissue identity is represented by colors and species by shape. The left panel shows all tissues, while the right panel highlights neural and testis samples compared to all others. Coordinate distributions of these three groups of meta-samples are shown on the side of the relative component. The percentage of variance explained by each PC is reported on the relative axis. **b**. Percentage of variance explained by the first 15 principal components from the PCA described in a. **c**. -log10(p-value) of two-sided ANOVA tests performed among the coordinates of the specified groups on each component. For the left panel (green) we tested if there was a significant difference between tissues or species groups. For the center and right panel (blue and orange) we tested if there was a significant difference between any query group (that is, column) versus all other collapsed groups. All tests were performed with the *aov* function in R, and p-values were Bonferroni corrected. **d**. Heatmap showing the clustering of tissues and species (rows) based on the expression across tissues of best-ancestral bilaterian-conserved orthogroups (columns). Expression values were z-scored across tissues of the same species in order to minimize the inter-species variability (see Supplementary Methods for the definition of the best-ancestral orthogroups z-scored expression matrix). Only the 2,436 best-ancestral orthogroups conserved in all species were considered. The heatmap was generated by the *pheatmap* function in R with *ward.D2* clustering method. Tissue colors refer to panel a.

https://doi.org/10.1038/s41559-024-02398-5

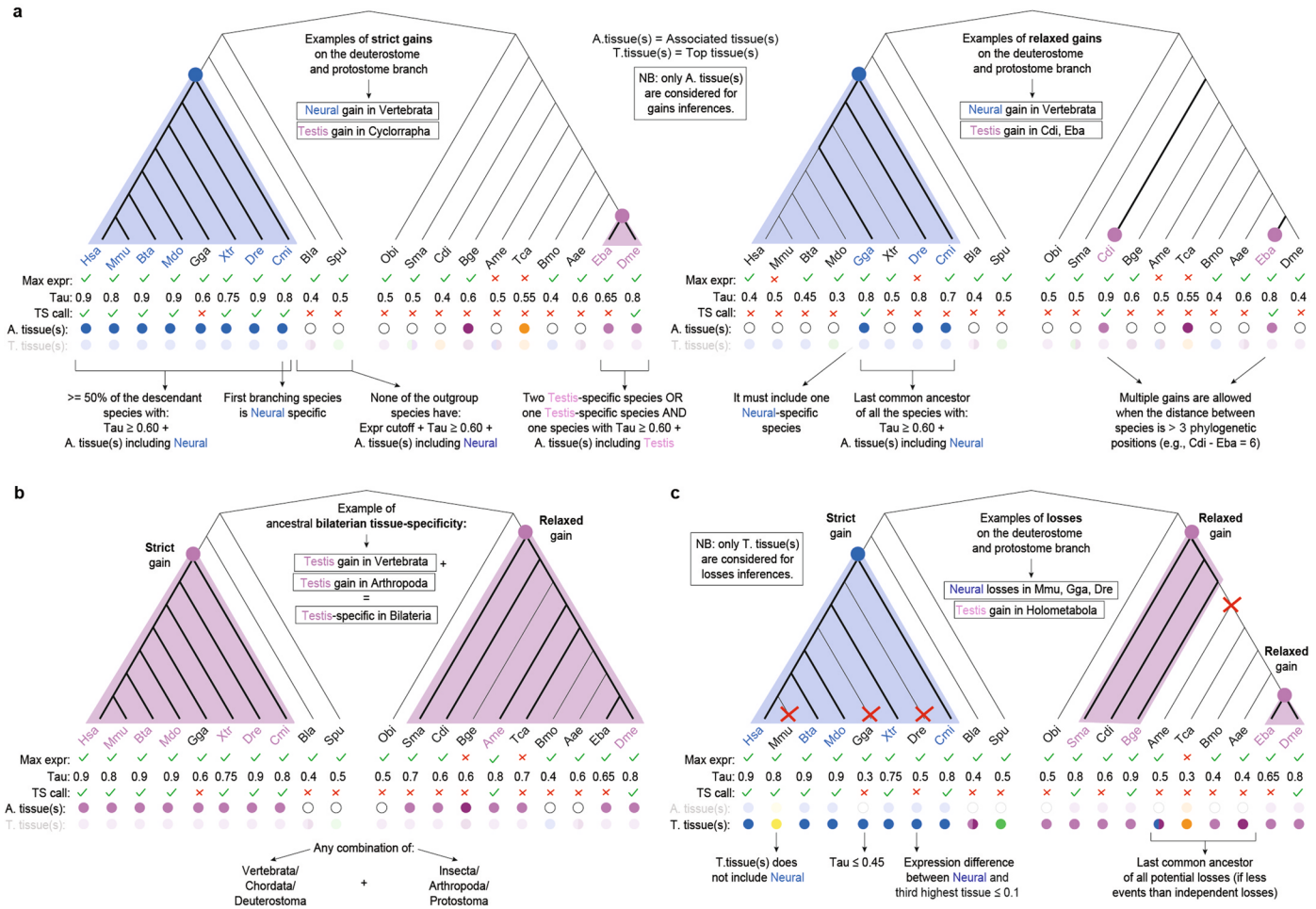


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41559-024-02398-5

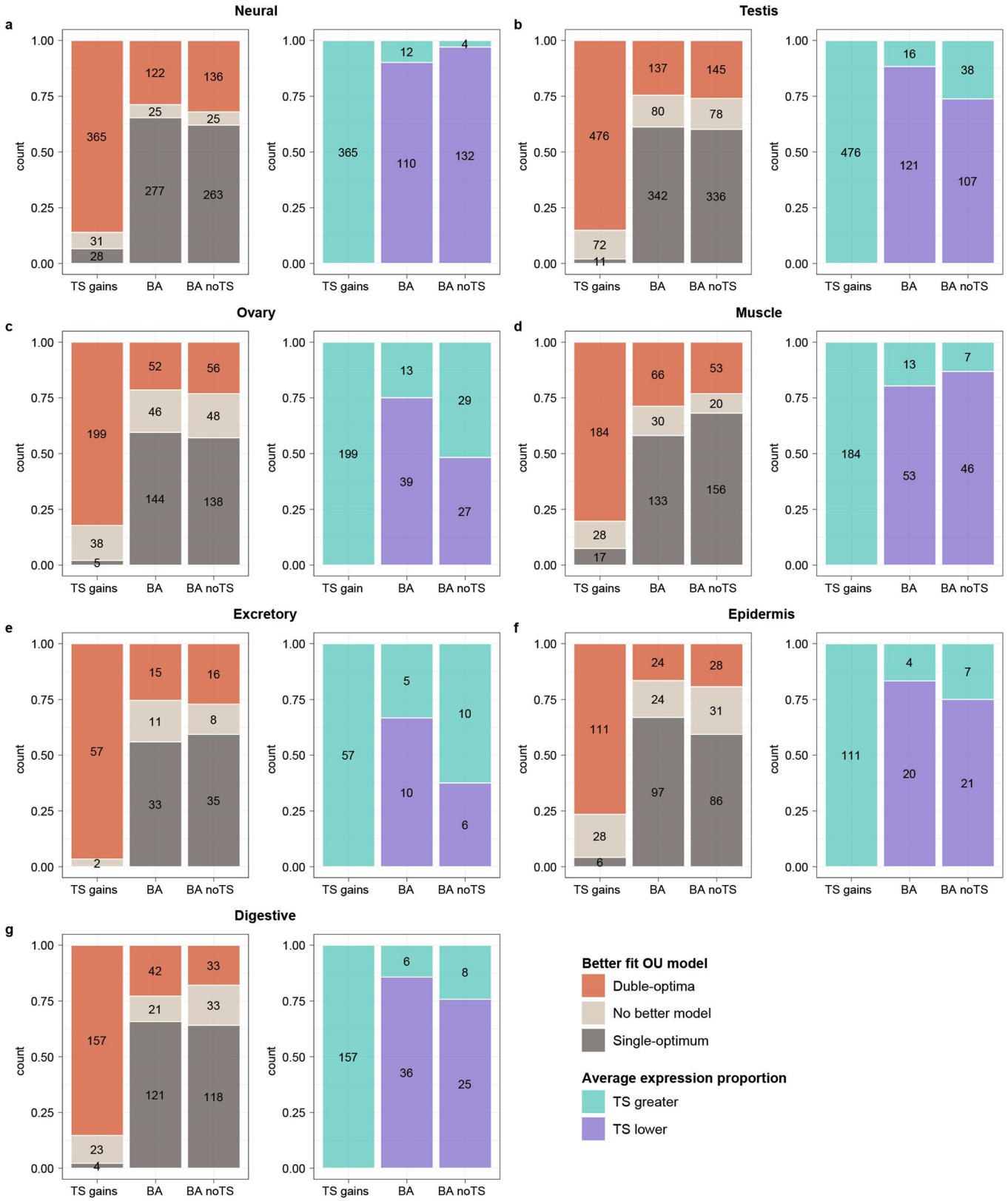
Extended Data Fig. 4 | Expression profiles of ancestral bilaterian tissue-specific expression modules. a-f. Coordinates of components returned by a sparse partial least square discriminant analysis (sPLS-DA) run separating the meta-samples of each tissue group (depicted with the relative colors) from all the others (grey). All 7,178 best-ancestral orthogroups were considered. The loadings of these components will be used to define the ancestral bilaterian tissue-specific modules (see Fig. 2a,b). The percentage of variance explained by each component is reported on the relative axis. **g-l**: Expression profiles across tissues of best-ancestral orthogroups in the ancestral tissue-specific modules (see Fig. 2c,d for neural and testis modules). (l) ovary module (n = 42); (h) muscle module (n = 112); (i) excretory module (n = 29); (j) epidermis module (n = 17); (k) digestive module (n = 51); (l) adipose module (n = 6). Expression values were first z-scored by species, and each dot represents the median expression among vertebrates, insects or outgroups. The boxplot features are defined as follows: the center line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend respectively to the lowest and highest points, to a limit of 1.5 multiplied by the interquartile range from the closest hinge.

https://doi.org/10.1038/s41559-024-02398-5



Extended Data Fig. 5 | Examples and criteria for phylogenetic inferences of tissue-specificity gains and losses. a. Examples and criteria for the inference of tissue-specificity gains on either the deuterostome or protostome branches with the strict approach (left panel) and the relaxed approach (right panel). **b**. Example and criteria for the inference of ancestral bilaterian tissue-specificity. **c**. Examples and criteria for the inference of tissue-specificity losses. NB: the best-TS orthogroups are the ones considered for all inferences of tissue-specificity gains and losses (see Methods and Extended Data Fig. 2c,d).

https://doi.org/10.1038/s41559-024-02398-5

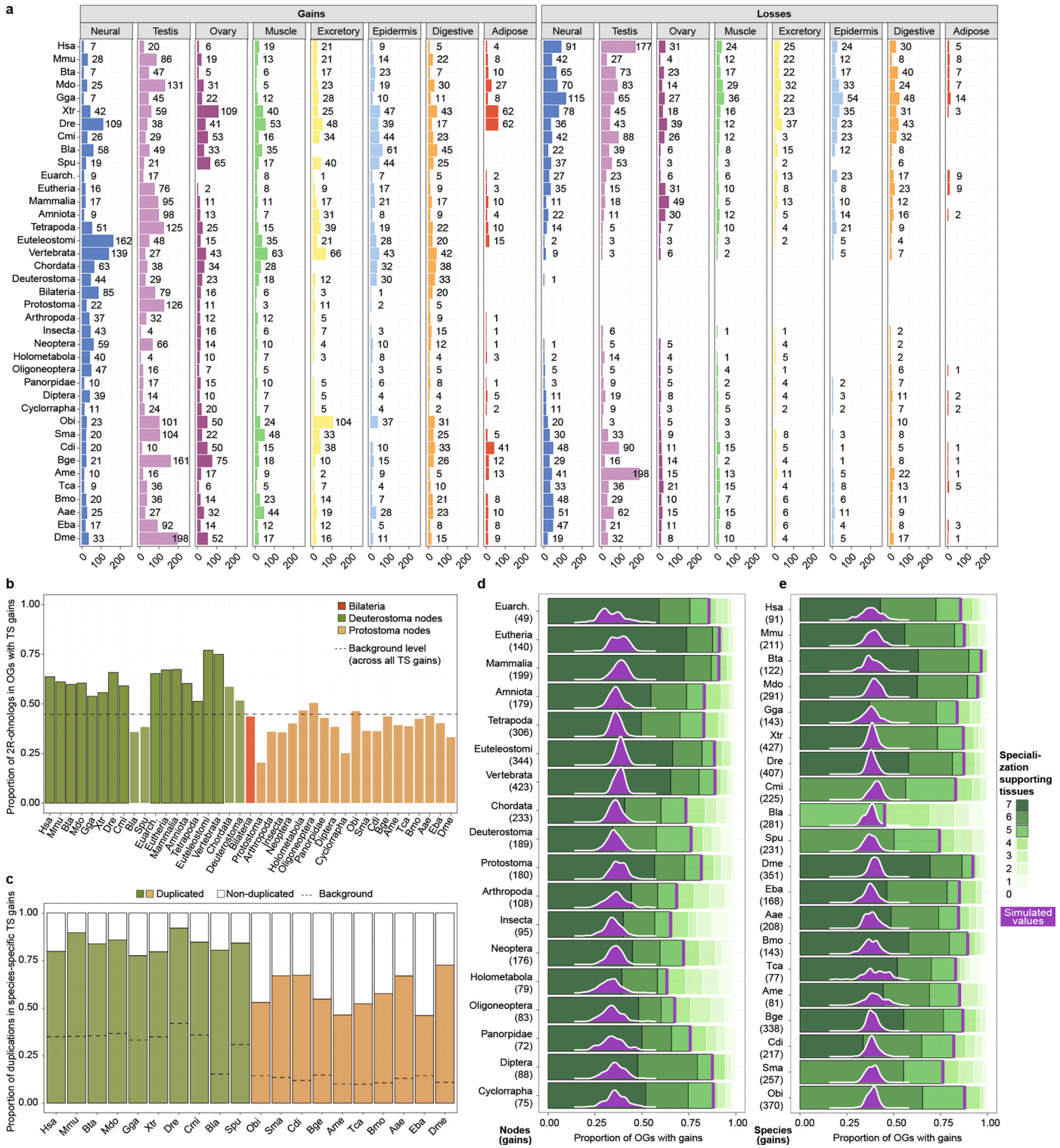


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41559-024-02398-5

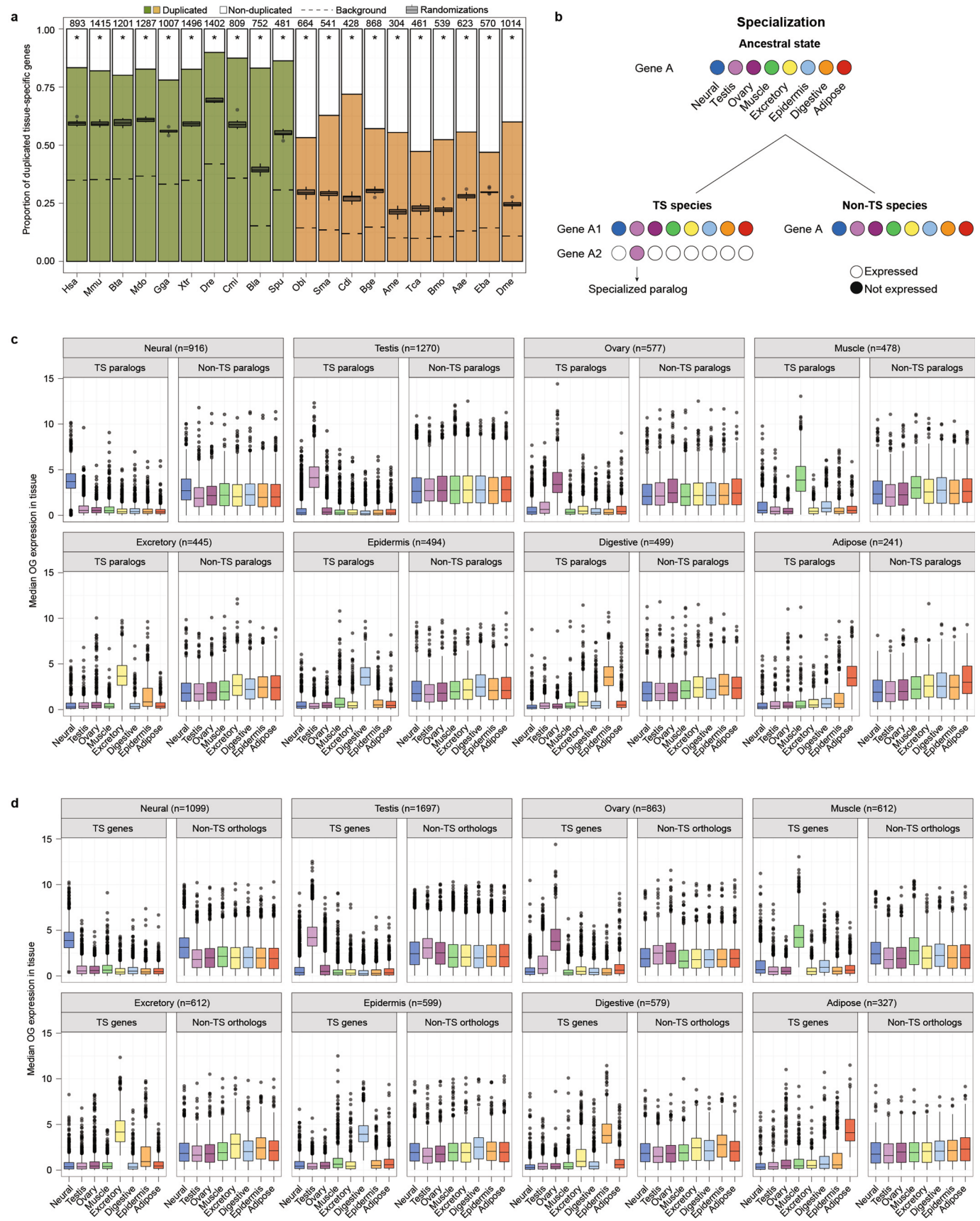
Extended Data Fig. 6 | Validation of inferred tissue-specificity gains. **a-g**. Orthogonal validation of all the inferred tissue-specificity gains in each tissue for which we could implement an OUs comparison method (see Supplementary Methods and Supplementary Discussion). The first bar always corresponds to the selected tissue-specificity gains (TS gains), while the second and third bars represent control sets (of the same size as the test set) sampled from either all best-ancestral orthogroups (BA) or best-ancestral orthogroups without tissue-specificity gains (BA no TS), to which we randomly assigned the tissue-specificity labels of the corresponding test set (see Methods). *Left barplot*: proportions of orthogroups based on the OU model (either a double-optima or a single-optimum) that better fits the relative expression levels. The double-optima OU model postulates different expression optima for the species with and without tissue-specificity, where the latter also include all species with losses. *Right barplot*: proportions of orthogroups better fitting a double-optima OU model (in red on the left barplot) depending on whether the species with tissue-specificity show higher/lower average relative expression compared to species without (TS greater/lower, respectively).

https://doi.org/10.1038/s41559-024-02398-5



Extended Data Fig. 7 | Extra statistics of tissue-specificity gains and losses. **a**. Barplots representing the number of inferred tissue-specificity gains (left) and losses (right) across all nodes/species (rows) and tissues (columns). Best-TS, bilaterian-conserved orthogroups were considered for these inferences. **b**. Proportion of tissue-specificity gains in each node/species occurring in best-TS orthogroups that include 2R-onhologs. Deuterostome nodes/species are distinguished between those diverging before (transparent color) or after (full color) the two rounds of vertebrate WGDs. The black line represents the proportion of 2R-onhologs across all tissue-specificity gains. **c**. Proportions of duplicated (that is, with at least one paralog) or non-duplicated (that is, single-copy) genes with tissue-specific, species-specific gains in all species. The background line represents the overall proportion of duplicated genes in each species. **d,e**. Same data represented in Fig. 4f, but plotted separately across all nodes (d) and species (e). **NB:** Bilaterian "gains" indicate ancestral bilaterian tissue-specificity, which might have been acquired either in the last bilaterian ancestor or previously in evolution. Abbreviations: Euarch: Euarchontoglires.

https://doi.org/10.1038/s41559-024-02398-5

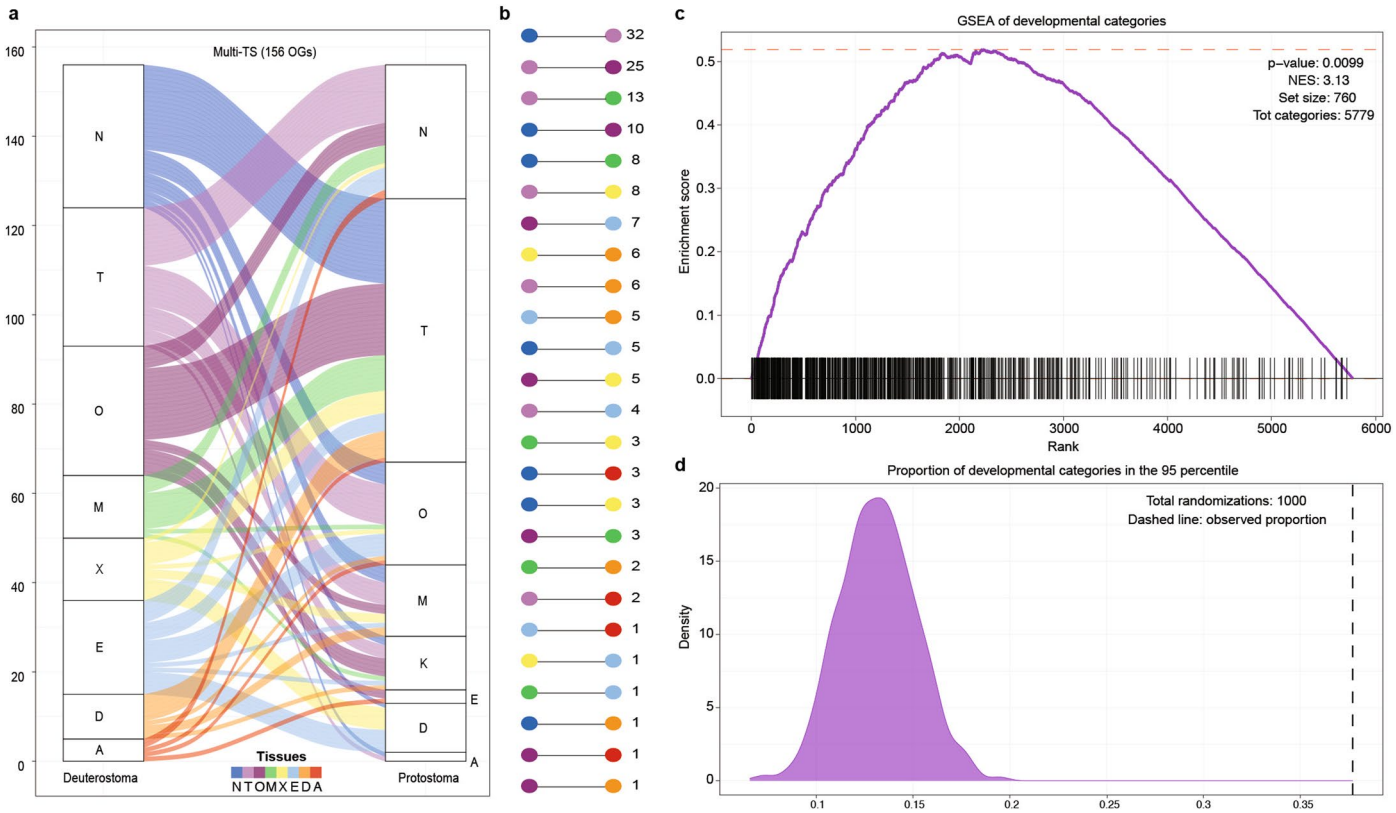


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41559-024-02398-5

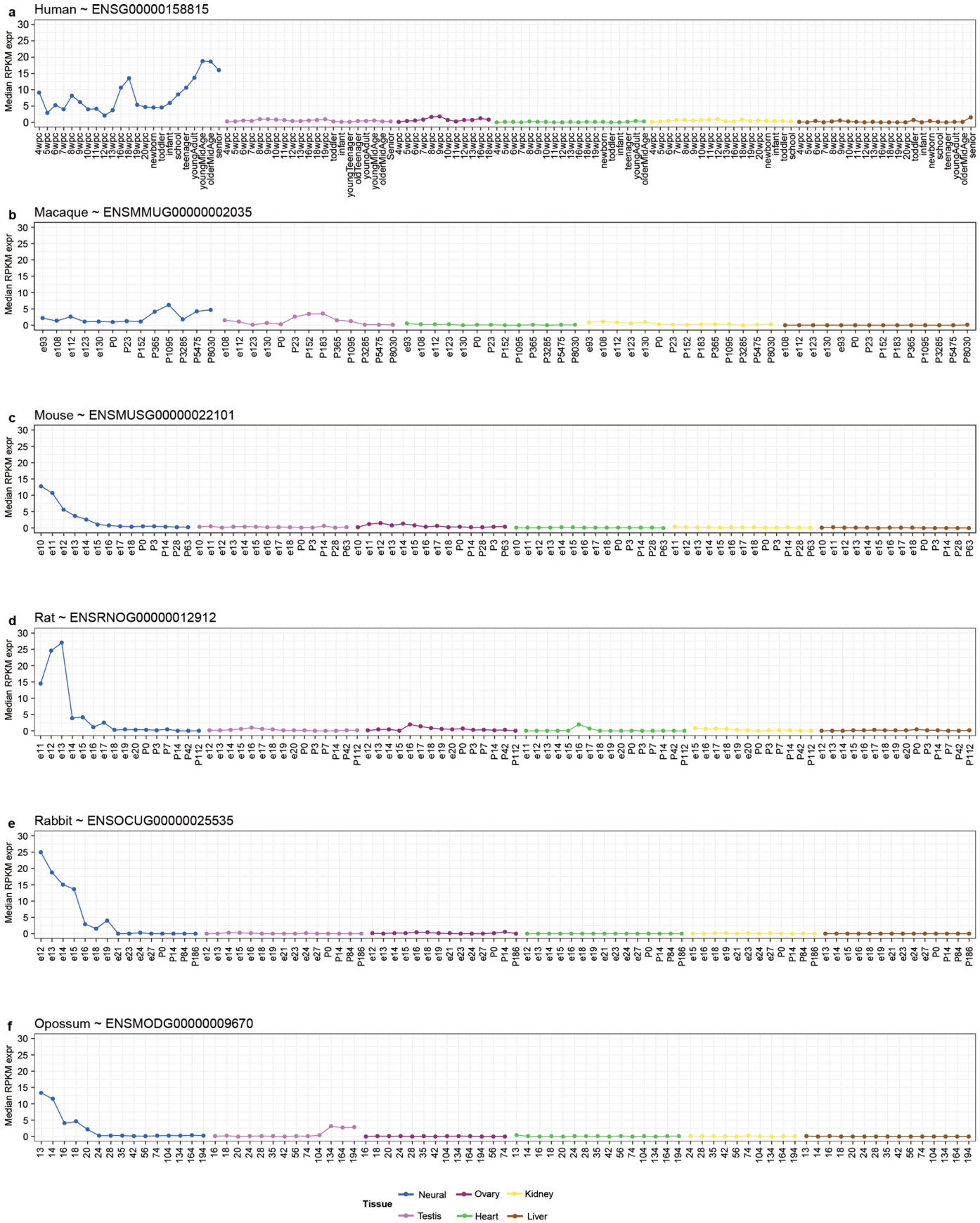
Extended Data Fig. 8 | Expression profiles of tissue-specific genes compared to non-tissue-specific orthologs and paralogs. a. Barplot: proportions of duplicated (that is, with at least one paralog) or non-duplicated (that is, single-copy) tissue-specific genes in each species. Boxplot: proportions of duplicated tissue-specific genes in each species upon the ten randomizations of the original orthogroups (see Methods). The asterisks indicate a significant difference (one-sided binomial test, alternative = "less"; p-value ≤ 0.05) between the observed proportion of duplicated tissue-specific genes and the median of such proportions coming from the randomization trials. The background line represents the overall proportion of duplicated genes in each species. The boxplot features are defined as follows: the center line represents the median; the lower and upper hinges correspond to the 25th and 75th percentiles; the lower and upper whiskers extend respectively to the lowest and highest points, to a limit of 1.5 multiplied by the interquartile range from the closest hinge. Outliers points are plotted individually. The total number of considered genes is reported above each species' bar. **b**. Scheme illustrating how tissue-specific expression can be gained following gene duplication and specialization. Color dots indicate expression in the relative tissue, white dots represent lack of expression. **c**. For each tissue, median gene expression in each bilaterian-conserved orthogroups for species possessing at least one tissue-specific and one non-tissue-specific gene. Expression of tissue-specific genes is plotted on the left, while expression of their non-tissue specific paralogs is shown on the right. Each data point in each tissue's boxplot is the median of the relative expression in that tissue for all corresponding genes and species. The total number of considered genes is reported in the relative plot. See panel a for description of boxplot features. **d**. Median gene expression across tissues for bilaterian-conserved orthogroups with tissue-specific gains in each tissue. Left: best-TS orthologs of the species with tissue-specificity. Right: best-TS orthologs in the other species. Each data point in each tissue's boxplot is the median of the relative expression in that tissue for all corresponding genes and species. See panel a for description of boxplot features. Distributions for gains within single nodes/species are available in the Supplementary Dataset.

https://doi.org/10.1038/s41559-024-02398-5



Extended Data Fig. 9 | Divergent and convergent evolution of tissue-specificity gains. a. Alluvia plot representing the best-TS, bilaterian-conserved orthogroups with tissue-specificity gains in distinct tissues between deuterostome (left) or protostome (right) nodes and species. Only orthogroups with gains in exclusively one tissue on each branch were considered. **b**. Number of parallel tissue-specificity gains between the deuterostome and protostome branch for all pairs of tissues represented in panel a. **c**. Plot from a Gene Set Enrichment Analysis (GSEA) testing for over-representation of developmental categories (760 out of 5779) among categories with high proportions of orthogroups that undergo species-specific gains of tissue-specificity. Only categories including at least 10 gene orthogroups were considered. The shown p-value refers to GSEA. **d**. Proportions of developmental GO categories among the top 5% (that is 95th percentile) of all GO categories ranked based on the proportions of their annotated orthogroups that undergo species-specific gains. The plotted values derive from 1000 randomization of the developmental labels among all GO categories, with the vertical dashed line corresponding to the observed proportion. Abbreviations: N: neural, T: testis, O: ovary, M: muscle, X: excretory, E: epidermis, D: digestive, A: adipose, NES: normalized enrichment score.

https://doi.org/10.1038/s41559-024-02398-5



Extended Data Fig. 10 | Developmental and adult expression of FGF17 in mammalian species. a-f: Expression values (RPKMs) for human FGF17 (a) and its orthologs in five mammalian species (b-f) across several developmental and adult timepoints in seven tissues. Data from[18].

nature portfolio

Corresponding author(s): Manuel Irimia

Last updated by author(s): Aug 10, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Public bulk RNA-seq data were downloaded from the NCBI Short Read Archive (SRA) through fastq-dump. All samples' metadata and mapping statistics are reported in Supplementary Table 3.
Data analysis	The analysis described in this paper were partly performed with public softwares (see below) and partly with custom pipelines (available on GitHub at https://github.com/fedemantica/bilaterian_GE). List of public software used in the study: - TransDecoder (v2.0.1) - stringtie (v2.1.3b) - hisat (v2.0.5) - gffread (v0.9) - broccoli (v1.2) - hisat (v2.0.5) - pigz (v2.3.4) - fastq-dump (v2.10.8) - fastqc (v0.11.5) - kallisto (v0.44.0) - DESeq2 (v1.22.2) - hashmap (v0.2.2) - tidyverse (v1.3.1) - tximportData (v1.10.0)

- tximport (v1.10.1)
- reshape2 (v1.4.4)
- stats (v3.5.2)
- pheatmap (v1.0.12)
- mixOmics (v6.6.2)
- gprofiler2 (v0.2.0)
- ggplot2 (v3.3.5)
- limma (v3.38.3)
- fgsea (v1.8.0)
- ape (v5.3)
- cowplot (v1.1.0)
- data.table (v1.12.8)
- ggalluvial (v0.12.2)
- viridis (v0.5.1)
- stringr (v1.4.0)
- ggtree (v1.14.6)
- ggpubr (v0.4.0)
- ggrepel (v0.8.2)
- ggridges (v0.5.2)
- gridExtra (v2.3)
- RColorBrewer (v1.1.2)
- forcats (v0.5.1)
- gage (v2.32.1)
- argparse (v1.1)
- pandas (v1.0.1)
- re (v2.2.1)
- numpy (v1.18.1)
- Bio (v1.76)
- scipy (v1.4.1)
- json (v2.0.9)
- ntpath
- itertools
- collections
- glob
- math
- subprocess
- sys
- os
- time
- urllib
- mafft (v7.222)
- revigo (online version, updated on march 22nd 2022)
- gimme_cluster
- rsat
- esearch (v15.6)
- pfamscan.pl (release 27)
- perl (v5.26.2)
- gtools (v3.8.2)
- dplyr (v1.0.7)
- ggforce (v0.3.2)
- lattice (v0.20-38)
- grid (base R package)
- stats (base R package)
- rstatix (v0.7.0)
- ggh4x (v0.1.2.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The FASTQ and processed files of the RNA-seq samples generated for this project are available at GEO under series GSE205498. The Supplementary Dataset is available at https://data.mendeley.com/drafts/22m3dwhzk6/2.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to pre-determine sample size. When multiple samples were available for a given tissue and species, we grouped the RNA-seq samples into a maximum of three meta-samples. This was done to: (i) increase read depth per meta-sample, (ii) dilute potential batch effects from publicly available samples, (iii) facilitate downstream analyses and comparisons by having a comparable number of replicates across tissues and species. Meta-sample groups are detailed in the column "Metasample" in Supplementary Table 3. In particular, we followed the approach that we previously described for human, mouse, cow, zebrafish and drosophila in VastDB [https://vastdb.crg.eu/], where samples from comparable experiments based on clustering approaches are pooled.
Data exclusions	Once the final dataset was assembled, no data was excluded from the analysis However, when several public RNA-seq samples for a given tissue and species were available, we discarded from the final dataset those samples showing poor similarity (in terms of gene expression correlation) with all the other samples from the same tissue and species.
Replication	Given the computational nature of this study, all results on our dataset can be reproduced with the scripts and datasets provided.
Randomization	RNA-seq samples were divided into groups (meta-samples) depending on the species and tissue of origin. However, we adopted one randomization-based control in order to test the robustness of our inferences of tissue-specificity gains and losses. In particular, we computed Tau values across species after randomizing the samples of each tissue across the relative meta-samples. Moreover, we used a randomization-based approach to test to what extent the observed association between gene duplication and tissue-specificity occurs by chance. To do that, we generated ten randomized bilaterian-conserved gene orthogroups by shuffling genes within each species while preserving the original paralogy structures (i.e., each randomized orthogroup conserved the original number of paralogs from each species, but the actual orthologous genes no longer corresponded across species). Finally, we adopted randomization-based strategies to test the incidence of specialization events (i.e., randomization of tissue-specificity labels within orthogroups) and the enrichment of developmental GO categories in species-specific, tissue-specific gains (i.e., randomization of GO labels).
Blinding	Given the computational nature of this study, blinding was not considered relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	We generated RNA-seq samples from tissues of 11 laboratory animals, which comprise 3 vertebrate and 8 non-vertebrate species. The selected species included opossum (Monodelphis domestica, a 39-week old female individual provided by the Turner lab at The Francis Crick Institute, London), frog (Xenopus tropicalis, an adult male individual by the Centro Andaluz de Biología de Desarrollo, Sevilla), zebrafish (Danio rerio, adult individuals from the Centre for Genomic Regulation, Barcelona), sea urchin (Strongylocentrotus purpuratus, adult individuals provided by Stazione Zoologica Anton Dohrn, Napoli), fruit fly (Drosophila melanogaster, adult individuals provided by the Department of Genetics, Universitat de Barcelona), hoverfly (Epysirphus balteatus, adults individuals provided by the Centre for Ecology and Conservation, University of Exter), yellow fever mosquito (Aedes aegypti, adult individuals from the Florida International University), silkworm (Bombyx mori, adult individuals provided by the Institute of Evolutionary Biology, Barcelona), red flour beetle (Tribolium castaneum, adult individuals provided by the Institute of Evolutionary Biology, Barcelona), cockcroach (Blattella germanica, adult individuals provided by the Institute of Evolutionary Biology, Barcelona), and mayfly (Cloeon dipterum, adult individuals provided by the Department of Genetics, Universitat de Barcelona).
Wild animals	We generated RNA-seq samples from tissues of 4 wild animal species, comprising 1 vertebrate (elephant shark) and 3 non-vertebrate species (octopus, centipede and amphioxus). Elephant shark (Callorhinchus milii) tissues were collected from a wild caught adult female captured by rod and reel fishing from Western Port Bay, Australia. Sex was determined by external morphological features (lack of claspers). The elephant shark was transported in a 4000L tank containing seawater to a 40,000L closed marine aquarium system, where it was temporarily housed until it was humanely killed for tissue collection as previously described (Boisvert et al 2015 Zoo Biology 34: 94-98). The elephant shark was humanely killed by immersion in the fish anaesthetic Aqui-S added to sea water, and targeted tissues collected into RNA Later and stored at -80C. For octopus, materials for sequencing were derived from wild-caught octopus bimaculoides supplied by Aquatic Research Consultants (Catalina Island, California, USA). Octopuses were anesthetized with 2% ethanol in seawater and euthanized by transecting the brain before tissues were isolated for sequencing. The centipede (Strigamia marittima) specimens were collected from a wild population near Brora, Scotland. Tissues from 15-20 individuals were dissected, preserved with RNA Later and transported in cold blocks to Barcelona, where the RNA was extracted. In the case of amphioxus (Branchiostoma lanceolatum), adult specimens were collected from the wild at the Racou beach near Argelès-sur-Mer, France, (latitude 42° 32′ 53″ N and longitude 3° 03′ 27″ E) with a specific permission delivered by the Prefect of Region Provence Alpes Côte d'Azur (as described in Marletaz et al, Nature, 2018. doi: 10.1038/s41586-018-0734-6).
Reporting on sex	Samples from reproductive organ tissues have been collected from individuals of the respective sex in each species. Samples from the other tissues generally include individuals of both sexes, unless otherwise specified above. For non-vertebrate species, tissue samples are derived from pools of individuals of undetermined sex.
Field-collected samples	The elephant shark was temporarily maintained in a 40,000L closed marine aquarium system as previously described (Boisvert et al 2015 Zoo Biology 34: 94-98) until it was humanely killed for tissue collection. The amphioxus unripe adults were maintained in filtered seawater during one week with one water change per day before RNA was extracted.
Ethics oversight	Elephant shark capture and tissue collection was approved by Monash Animal Ethics Committee ERM 14162. All work on octopus was performed in compliance with the EU Directive 2010/63/EU on cephalopod use and AAALAC guidelines on the care and welfare of cephalopods; the study protocol was approved by the Institutional Animal Care and Use Committee at the Marine Biological Laboratory. Zebrafish procedures were approved by the Barcelona Biomedical Research Park Institutional Animal Care and Use Ethic Committee (PRBB–IACUEC). All experiments involving frogs conform to the national and European Community standards for the use of animals in experimentation and were approved by the Ethical Committees from the University Pablo de Olavide, Consejo Superior de Investigaciones Científicas (CSIC), and the Andalusian government, under the project nº 03/05/2018/065. Opossum procedures were approved by the Francis Crick Institute and the UK Home Office.

Note that full information on the approval of the study protocol must also be provided in the manuscript.